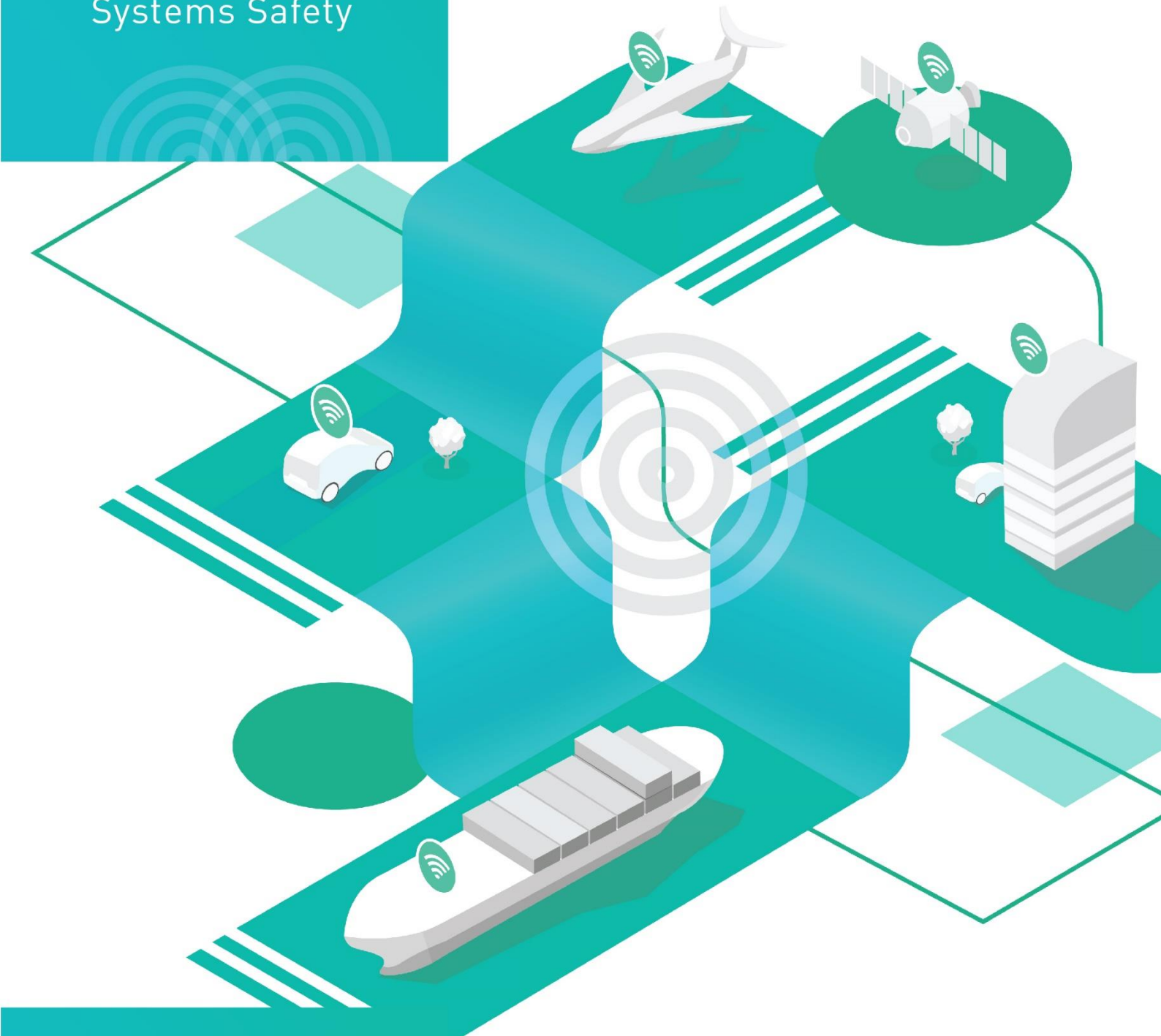
IWASS

First International Workshop on Autonomous Systems Safety

Proceedings



11th to 13th March, 2019
Trondheim | Norway

NTNU
Norwegian University of Science and Technology

The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING

IWASS

Proceedings of the First International Workshop on Autonomous Systems Safety

Edited by:

Marilia A. Ramos, Christoph Thieme, Ingrid B. Utne, Ali Mosleh

NTNU
Norwegian University of Science and Technology

The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING

Preface

The First International Workshop on Autonomous Systems Safety (IWASS) was organized by the Department of Marine Technology at the Norwegian University of Science and Technology (NTNU) and the B. John Garrick Institute for the Risk Sciences at the University of California, Los Angeles (UCLA). IWASS took place in Trondheim, Norway, from 11th to 13th of March 2019. The participation in the workshop was by invitation only and included 47 subject matter experts from Europe, Asia, Australia, and the U.S., working in both academia and industry.

The idea to organize IWASS originated from a discussion by the organizers on the challenges concerning autonomous ships' safety. A natural question that emerged was whether such challenges are unique to the maritime domain or there are commonalities with other technology sectors such as aviation and land transport. And if there are identifiable similarities, could common solutions be envisioned and developed? Answering these and related questions motivated IWASS as a platform for an interdisciplinary discussion on risks, challenges, and foremost potential solutions concerning safe autonomous systems and operations.

The awareness on autonomous systems' similarities is not a novelty in the field. Yet, to our knowledge, no event before IWASS had assembled experts on different autonomous systems with the purpose to discuss safety, reliability, and security (SRS). In the past, similar events have been organized around a specific type of autonomous system (e.g. cars, ships, aviation) or a particular safety or security related aspect of such systems (e.g., the risk of cyber-attack, software reliability). IWASS distinguishes itself from these events – and complements them – by bringing these topics together in an attempt to focus on proposing solutions for SRS challenges common to different types of autonomous systems.

These proceedings document the discussions and the main results from the workshop. It includes (i) an overview of the different applications of autonomous systems and its challenges concerning SRS; (ii) summaries of the discussions on key topics held at IWASS; and (iii) complementary papers from several participants.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

Introduction

The introduction of automation is rapidly changing how society interacts with machines. The enablers include new applications of artificial intelligence, machine-learning, and powerful integrated software and hardware systems. The push towards higher levels of autonomy is challenging how society should safely design, operate, interact, approve, and accept such systems. The challenges are interdisciplinary in nature and require a collective effort by the research communities to categorize and classify risks and conceptualize and develop mitigating solutions. Hence, the workshop aimed at discussing challenges related to safety, reliability, and security (SRS) of autonomous systems, covering autonomous maritime, marine, land vehicle, railway, and aerospace systems, and proposing possible solutions to the identified challenges.

A Whitepaper prepared by the organizers and sent to the participants before the workshop presented preliminary challenges regarding SRS of autonomous systems. The whitepaper provided a background for the discussions of the workshop and the interdisciplinarity of the topics, namely: (i) Interaction of software, hardware, and human operators; (ii) assessment methods for safety, reliability and security; (iii) cyber security; (iv) legal and regulatory aspects; and (v) ethical and social aspects.

The topics above guided the presentations and discussions held during IWASS. The first day of the workshop was dedicated to lectures by experts from both the industry and academia. The summaries of these lectures are presented in these proceedings, following the take-away messages. The lectures covered topics ranging from cybersecurity to trust in autonomy and, along with the whitepaper, provided a foundation for discussion in four breakout sessions during the second day of the workshop. During the breakout sessions the participants addressed challenges and possible solutions related to (i) human factors, ethics, and regulations, (ii) SRS methods and modeling, (iii) system verification and testing, (iv) intelligence and decision making. The discussions were guided by a proposed setoff research questions from Table 1. Preliminary conclusions of the breakout groups were presented and discussed during the final day of the workshop. Each group further prepared a report of their discussions, which have become chapters in these proceedings. In addition, the participants were invited to submit their recent work on autonomous systems' SRS.

These proceedings document the discussions and the main results from the workshop. Following this introduction, the whitepaper presents an overview of the state of the art of autonomous systems development and related SRS

challenges. The following Chapters provide summaries of the four breakout group discussions, namely: 1) Autonomous Transportation Technologies: Society and Individuals in the Loop; 2) Safety, Reliability and Security Modelling and Methods for Autonomous Systems; 3) System Verification, Processes and Testing; and 4) Intelligence and Decision Support. Thereafter follows papers from several participants addressing topics from the workshop.

Take-away Messages

Autonomous systems are emerging and are crucial for enabling new operations, such as autonomous land-based, maritime, and air transportation, mapping and monitoring of oceans and areas on land, and inspections of physical structures that are difficult to access. Autonomous technologies are intended to be a step towards safer and efficient operations, but the corresponding software and advanced control systems also involve complexities that pose formidable challenges to identification and removal of cause of functional failure, safety issues, and security concerns.

Developers of autonomous systems and autonomous functionality must develop strategies and methods for safe, reliable and secure performance so that the autonomous systems comply with requirements. *Operators* need to plan and execute safe and robust operations, with effective risk control. The *authorities* need new standards and guidelines for autonomy as a basis for their regulatory and oversight activities. Acceptance and approval of autonomous systems with integrated learning and optimization capabilities require *risk identification, assessment, modelling, testing and verification as drivers of the design, operation, and system validation*.

Widespread acceptance of any new technology depends on *trust* in the technology itself and in the organizations that implement and regulate it. This means that the technology itself must not produce significant unpleasant surprises and the organizations must demonstrate competence and act consistently with the societal values. Hence, *social acceptance* of autonomous systems, composed of advanced control, decision, and sensor systems, with increased intelligence, requires open and inclusive development processes to determine what societal *values* are embedded and affected by the technology.

Autonomy in systems adds new and formidable complexity with respect to ensuring SRS. The main challenges related to effective and adequate methods for SRS assessment concern the *inadequacy of existing methods*, *the lack of integrated modelling of hardware, human, software*, the challenge of *learning systems*, and *data requirements*.

Software malfunctioning, and cyber threats are different from hardware and human failures. Past failures do not indicate future behavior which means that calculation of the expected likelihood or frequency is not feasible. In addition, learning capabilities of the software increase the difficulty in validating performance. Nevertheless, many of the current methods can still play a part in supporting SRS of autonomous systems, but *new modelling techniques*, which holistically capture the strong connectivity and interdependencies between software, hardware and human operators, are required. Further, *simulations* may assist in the detailed understanding of autonomous systems behavior, identification of SRS issues, and performing system validation. The importance of quantifying SRS may increase in the future to enable *real-time decision making* and to identify when the system performance drops below the acceptable threshold during operation.

The *verification and testing* of autonomous systems are related to *regulatory, societal, and ethical requirements*. These requirements need to be included at the beginning of the design process and fed through the verification phase. A main concern, however, is how to obtain the necessary requirements against which to verify the system. Validation of requirements is a concern for the verification of any system, but it may be a particular challenge with autonomous systems because of the *complexity* and a lack of consensus on regulation and ethical guidelines for autonomous systems.

Further, the identification of the best *verification processes* for autonomous systems is a concern. Due to the complexity of the such systems, and their potential for learning, continuous and integrated processes seem to be needed. *Communicating* the verification efforts to regulators and the public is important to ensure that the autonomous systems can be certified by a regulator and trusted by the public. *Formal methods* can provide automatic verification and unambiguous specification of the autonomous system's intended behavior.

Autonomous systems are *cyber physical systems* and may include *Artificial Intelligence (AI)*, for example, in their automated decision making. The capabilities and interpretation of AI is often vague and misunderstood by the public. AI is a collection of mathematical methods, helpful for solving tasks associated with intelligence. AI is mainly comprised of some form of learning, some degree of reasoning, interaction with an environment, and should have the ability to explain how or why the AI makes its internal decisions. Reasoning is an important aspect of autonomous systems and AI and refers to the ability to explain actions and decisions.

AI methods try to find regularities in sets of data. AI has some advantages over humans, for example, enabling quick analysis of large sets of data. Therefore,

AI may be used to learn the *operational parameters* for an autonomous system, *identify weights for risk factors*, or to detect system or operational *deviations*.

Summaries of lectures and the discussions in the breakout sessions are provided in the following sections.

Summary of the Lectures[1]

Modelling and assessing risks of autonomous systems: Challenges and perspective on solutions

Ali Mosleh, The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles, USA

This talk characterized autonomous systems as Cyber Physical Human (CPH) systems, noted by their complexity, heterogeneity, functional and physical distribution, interconnectivity of technology and social dimensions, and openness, and learning ability. The development and deployment of CPH occurs with a high pace through often distributed supply chains of varying quality, reliability, and safety standards. All these attributes may lead to emergent failures that are rare and difficult to identify, with possibly catastrophic consequences, often through conflicts and masking effects.

The presentation argued that traditional modeling and analysis methods (such as FMEA and FT) have significant limitations in analyzing and improving safety, reliability, and security (SRS) of CPH systems in general and autonomous CPH systems in particular. The talk suggested that a very promising approach is dynamic probabilistic simulation of such systems. An overview of the emerging simulation methods for risk analysis was provided.

The Norwegian maritime authority´s approval process of autonomous ships - Our challenges and guideline

Nils Haktor Bua, Norwegian Maritime Directorate, Norway

When it comes to new technology and alternative design the Norwegian Maritime Authority (NMA), as the flag authority, is the one to evaluate the projects towards a certificate. As a basis for evaluating alternative design, NMA uses IMO circ. 1455, which gives the general process. For projects regarding autonomy and in which people are taken away from functions onboard ships, NMA is working on a guidance for the evaluation process, with basis in IMO circ. 1455.

Any alternative design needs to be shown to be as safe or safer than a conventional design. The burden of proof for showing this lays on the project. The

[1] The presentations held at IWASS are available and can be downloaded at: https://www.ntnu.edu/web/imt/iwass/presentations

presentation pointed out this process and what needs to be evaluated and shown during such a process. At the end, some of the challenges with projects with new technology and autonomy were listed.

Qualification of autonomy for risk and regulation - A behavioral approach

Tristan Perez, Boeing Research and Technology, Australia

Trusted operations of Autonomous Systems (AS) with increased levels of autonomy, moving from remotely operated systems onto higher levels of system decision making and autonomy to act, are far from being enabled today. Enabling these systems and their operations will require the development of trust in the combined technological, regulatory, and social environments.

This presentation discussed some technological challenges associated with autonomous system capability and a potential framework to assess system behaviours in relation to risk assessment and certification. The approach takes a behavioural viewpoint of mathematical system theory, links it to an epistemic view of uncertainty quantification and finally to the decision-making process of different stake holders. Moreover, the presentation reviewed part of the previous and current work at Boeing in these areas and discuss challenges and extensions that will be needed to assess systems with high levels of autonomy.

Industry perspective on the development of autonomous busses - Robustness development

Matthew Minxiang Hu, Haylion Technology, China

This presentation gave an overview of the efforts undertaken to design and test a robust autonomous bus system. The main consideration when designing a robust system is to make the product/process insensitive to uncontrollable user environments. This includes the users' interaction with the system and the manufacturing process, which may deviate from the original intend. Five main strategies can be employed to make a system more robust: (i) change the design concept, (ii) use design assumptions that are insensitive to the noise, (iii) reduce and remove noise factors, (iv) insert a compensation device, and (v) disguise the noise effects.

Unmanned aerial systems and risk

Adrian Arjornilla, UAS consulting, USA

This presentation summarized ongoing research work in the UAS CG and challenges to be dealt with when operating unmanned aerial systems from an operator perspective. Human workload is seen a key factor regarding the level of risk of an operation. The workload is driven by mission complexity, aircraft status, communication, and environmental conditions. In a current research project, it is attempted to integrate and display information intelligently, such that the workload is optimized, and the situational awareness is maximized. Important questions that need to be answered for such a system are: How can the handover between autonomous flight and manual flight be assessed in terms of risk and used as decision criterion for handover? How can the risk be measured in real time and predicted in the future? How can risk models be used to mitigate risk proactively? How can an autonomous system be assessed for their airmanship (similar to pilots)? How to make risk informed decisions during normal operation and emergency situations?

Cybersecurity for autonomous systems – Vulnerabilities and threats

Kenneth Titlestad, Sopra Steria, Norway

To assess safety, reliability and security for autonomous systems we primarily consider the three factors: software, hardware and human-in-the-loop. For properly addressing cybersecurity for such systems we should also take into account the possible cyberattack-agents which act as ghosts-in-the-loop. In the presentation this was exemplified with the Trisis-attack at a Saudi-Arabian petrochemical plant in 2017. At the plant there were several, repetitive malfunctions of a specific type of safety-controller. The safety controllers are part of the most critical automation systems at the plant, and they are in place only to detect unsafe conditions in the production process. When this is detected these controllers automatically run production shutdown or emergency shutdown. At the specific plant in Saudi-Arabia, after weeks of troubleshooting, including malfunctioning of several replacement units, cybersecurity-analysts detected an advanced, unknown malware in the safety-networks. This malware exploited vulnerabilities of the controllers and replaced the firmware on them once they were installed in the safety-network.

All autonomous systems have central controllers that act as the hardware- and software-based brains of the system. These get input from sensors and provides output to effect-generators, such as actuators and motors. The controllers also provide an interface into the system for the Human-Machine-

Interface (HMI). All parts of these systems are increasingly consisting of programmable components, which results in new unknown vulnerabilities, and new unknown ways of attacking the systems. This problem is accentuated in Operation Technology (OT) and autonomous systems due to a higher focus on availability compared to systems in Information Technology. Doing security upgrades on the former is a rigid process that cannot be done often, which results in most parts of the systems on a daily basis having more security vulnerabilities than in an upgraded state.

In the presentation several known attack vectors were presented and some of the major cyberattacks within industrial systems are listed. Trends and possible future threats were discussed. The presentation concluded with an open challenge to the industries to establish barriers which provide safety more by the laws of mechanics, physics and electronics. This is already being done within electronic opto-isolators, rupture discs, spring-return-valves and can be further developed for industrial- and autonomous systems for example with the use of data diodes to segregate safety-critical components from the rest of the system.

Intelligent machinery systems for autonomous ships

Sverre Torben, Rolls Royce Marine, Norway

Rolls Royce Marine delivers services ranging from ship design, over system integration to through-life support. Current efforts aim at making ships and their system more intelligent and digitize systems that are currently to a large extend analogous. Future ships systems need intelligent data collection to improve maintenance planning and situation awareness. Digital twins and platform clouds will enable shipping companies to make better design and operational choices.

Developments that are necessary for future remote and autonomous vessels are intelligent engine health monitoring, intelligent awareness systems, autonomous navigation system, all-speed track pilot system, and remote-control stations.

Trust in autonomy: Cyber-human learning loops

Asun Lera St. Clair, DNV GL, Norway

This presentation addressed the ethical and societal implication of autonomous systems. It argued that these are far more complex that the mere aspiration to embed ethical reasoning into algorithms. The presentation argued

that morality is a characteristic of human beings and cannot be transported into machines. It is important to distinguish between explainability of autonomous systems versus trustworthiness. Trust is underpinned by shared ethical and societal values, and the conditions for trusting technologies are similar to those of trusting other people or institutions. In the case of autonomy this means both, an assessment of the goals and purpose of the technology as well as assessment of the technical robustness of the system. The core ethical and societal issues associated with autonomous systems emerges from the complex interactions between software, hardware and human beings, alongside the context in which the system operates and the consequences it may have-- directly or indirectly, on people and the environment. Even if autonomous, human beings are part of their design, construction, deployment, operation, maintenance, evaluation and verification of these systems. A potentially normative approach to aim towards is the generation of cyber (physical)-human (social) learning loops, requiring true interdisciplinarity, in particular with the social sciences and the humanities.

Some recent advances in human-automation interaction design methods and future research directions for safety

David Kaber, North Carolina University, USA

This presentation reviewed recent advances in human-automation interaction modeling approaches, including new ideas to account for how tasks are interactively managed and traded by humans and machines. Another aspect of the work is focused on how these new design methods may be synergized and applied throughout the systems design and engineering cycle to better support human-machine system design. An additional section focused on the development of advanced vehicle automation based on current practices of automation design and implications for systems safety. This research reveals a paradox of automation for safety in which operator reliance on low-level automation for low severity hazard exposures may lead to skill decay for manual performance necessary in system negotiation of complex and high severity hazards.

Game theoretic simulation for verification and validation of autonomous vehicles

Anouck Girard, University of Michigan, USA

Autonomous vehicles have been the subject of increased interest in recent years in defense, industry and academia. Serious efforts are being pursued to

address legal, technical, and logistical problems and make autonomous vehicles a viable option for broad ranges of applications. One significant challenge is the time and effort required for the verification and validation of the decision and control algorithms employed in these vehicles to ensure a safe and reliable experience.

For example, for driving, hundreds of thousands of miles of tests are required to achieve a well calibrated control system that is capable of operating an autonomous vehicle in an uncertain traffic environment where interactions among multiple drivers and vehicles occur simultaneously. Traffic simulators where these interactions can be modeled and represented with reasonable fidelity can help to decrease the time and effort necessary for the development of the autonomous driving control algorithms by providing a venue where acceptable initial control calibrations can be achieved quickly and safely before actual road tests.

In this talk, a game theoretic traffic model was presented, that can be used to model human-driven and autonomous vehicles and their interactions, test and compare various autonomous vehicle decision and control systems and calibrate the parameters of an existing control system. The simulator is highly scalable and can handle several dozen interacting vehicles in near real time. The presentation demonstrated applications to highway driving and intersections, and discussed extensions to other transportation domains.

Breakout sessions – Research Questions

The breakout sessions addressed the research questions in Figure 1, which provided the overall scope of the discussions. Chapters 1-4 give more insights into the details.

Table 1: Research Questions

General research questions:
What are currently the main challenges with respect to safety, reliability, and security (SRS) of autonomous systems and operations?
What are similar challenges between the different autonomous systems, operations and industries?
Are there challenges with respect to SRS that only apply to one or few applications (and not all)? If so, which?
Which type of autonomous system/ operation is most advanced currently (has the highest level of autonomy and highest complexity)?
Which type of autonomous system/ operation is most feasible to realize in the near future? Why?
Human factors, ethics, and regulations:
How can we make risk related autonomous systems acceptable for society for widespread use?
Monitoring, remote operation and supervision from some sort of control center is relevant for many autonomous systems. This means that huge amounts of data are generated, collected, and stored. In an accident investigation, how can the authorities ensure that official investigation groups are authorized correct access to the data they need to identify root causes?
Who is responsible for decisions made by an autonomous system (in case of accidents)?
How should autonomous systems communicate in operation? How should they communicate with non-autonomous systems?
How should autonomous systems communicate safety to users and third-party stakeholders?
Do autonomous systems and operations need to be "as safe as" or safer than other types of systems?
Are AI systems suitable to make ethical decisions, what would be necessary for ethical decisions?
SRS methods and modeling:
Risk assessment needs to be integrated in the early stage of the design and development phases of all kinds of technological systems. Is this more or less challenging for autonomous systems?
A challenge with hazard identification and risk analysis of novel systems is to identify everything that can go wrong. How can we deal with the unknown unknowns?

What type of data is needed to analyze safety and reliability of autonomous systems and control risks during operations? How should such data be collected and utilized?

How can risk assessments and risk models of autonomous systems take shared control and "adaptive autonomy" sufficiently into account in the identification of hazards and the analysis of risk?

How can risk analysis contribute to improved situation awareness and intelligence in autonomous systems?

How can we determine "acceptable risk" for autonomous systems and operations? Should "acceptable risk" change with level of autonomy (LoA)?

A holistic approach is needed for SRS assessment of autonomous systems. What does this actually mean?

How can vulnerabilities in the software and communication systems of autonomous systems be reduced to mitigate cyber-attacks and security problems?

System verification and testing:

How can verification and the corresponding test scope for autonomous systems be managed dynamically?

What role does standardization play in the realization of autonomous systems?

Online operational data can be useful for managing risks related to autonomous systems. How can we assess the system performance and correctness from such type of data?

Can improved health monitoring of autonomous systems justify less reliable components?

Intelligence and decision making:

Uncertainty in sensor data is a challenge. How should this uncertainty be handled in the design and operation of autonomous systems and operations?

Sensor performance is affected by weather and climate conditions. Are autonomous driving/ navigation systems, on land and at sea, possible to realize in areas with such environmental conditions as in Norway? How?

Improved intelligence and online decision-making capabilities are needed in autonomous systems. Existing control theoretic approaches are not explicitly connected to risk assessment and modeling. What parameters, constraints and cost functions should be developed for control algorithms to minimize risk?

The control system architecture can roughly be divided into three levels; the execution layer, the guidance and optimization layer and the supervisory layer. What risk reduction measures are needed in these layers?

Some reports and documents state that artificial intelligence (AI) is a threat to human existence? Is this true? If yes/ no – why and how?

How can risk modeling of hazardous scenarios be used for autonomous optimization-based decision making and control under uncertainty?

Revised Whitepaper of the First International Workshop on Autonomous Systems Safety

Autonomous Systems Safety – State of the Art and Challenges

Marilia A. Ramos[1]; Christoph Thieme[1]; Ingrid B. Utne[1]; Ali Mosleh[1,2]

[1]Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[2]The B. John Garrick Institute for the Risk Sciences, University of California in Los Angeles (UCLA), Los Angeles, U.S.A.

This whitepaper provided a starting point concerning some of the topics that were to be addressed at IWASS, including the current state of the art on autonomous systems development and challenges it faces. In the following pages, we discuss challenges in respect to risk assessment techniques, human-machine interaction, cyber security, regulatory issues, and ethical aspects.

Autonomy and autonomous systems

The introduction of automation in a wide range of activities has changed how society interacts with machines. For years, automation was applied only to physical activities, rather than cognitive aspects, such as, situation assessment, sense-making and decision-taking. The advent of artificial intelligence, machine-learning, and easier access to powerful software and sophisticated hardware have brought a new revolution into how we interact with automated systems, both as users as well as operators. The outcome of this revolution are highly automated and autonomous systems.

Autonomy can be defined as a system's ability to make independent decisions and to adapt to new circumstances to achieve an overall goal. This is achieved without additional input from human operators or other systems [1]. Automation, on the other hand, is often understood as the reproduction of an action, without any choice made by the machine executing the action [2]. The degree of autonomy of a system may be assessed through Level of Autonomy (LoA). Several authors have proposed different scales for LoA [3], either generalizable to autonomous systems or specific to an industry [1]. In general, the LoA scale starts at a lower level autonomy in which information reception from the system and surroundings, situation assessment, decision-making, and command giving to the hardware are responsibilities of human operators. The LoA scale progresses to a higher level, when these tasks become responsibilities

of a software. Between the lower and higher levels, these tasks are shared between software and human, as illustrated at the Figure below.

A system may be designed with an adaptive autonomy [4], or dynamic autonomy [5], i.e., it may operate as highly autonomous during part of its operation or for performing certain tasks, and then operate in a lower autonomy level for other types of operations. An autonomous system may also be both manned and unmanned.

Many areas of life and business comprehend systems with some level of autonomy. For instance, autonomous chatbots are found on the internet, autonomous manufacturing systems are taking up production, and autonomous transportation systems are being tested on land, in water, and in the air. Although the first industrial sectors to introduce some level of autonomy into transportation were aeronautics and the aero-spatial domains, significant investments have recently fast-tracked the development of autonomous cars, and put those in the spotlight.

The rapid evolution of technology enabling autonomous cars can be illustrated by the Grand Challenge, an event organized by DARPA[2]. The Grand Challenge consisted of a competition of autonomous cars to go through California's Mojave Desert. In 2004, no car finished the race and the most successful one, the Red Team's vehicle, reached a maximum of seven miles of the course. In the following year, five vehicles completed the race. In fact, Google's first project on autonomous cars was launched in 2009 with a team from DARPA veterans. The development of autonomous cars is driven today by giants of the

[2] DARPA (Defense Advanced Research Projects Agency) is an agency of the United States Department of Defense.

tech and auto industry, such as Google and Tesla, Ford, and General Motors. These are followed [3]by smaller startups as May mobility and Drive.ai.

DARPA Challenge 2004 Red Team's car (left side) and Waymo (formerly the Google self-driving car project) autonomous car (right side) [3]

Autonomy is also applied in other land transportation systems, such as buses and trains. China has launched the world's first self-driving bus in August 2015. The bus drives with guidance from cameras, lidars, and a master controller, along with a human driver behind the wheel, who should take over control in case of any problems. Other examples include the Norwegian city Stavanger, where the mass-transit company is testing autonomous buses, and Catalonia, Spain, where an autonomous bus called Èrica is being tested to help citizens become familiar with driverless technology. In Finland, three cities are expected to receive autonomous buses by 2020. The technology will be provided by the Japanese company Muji, and it should be the first autonomous bus in the world suited to all types of weather.

Land transportation on railways has also advanced using automated and autonomous systems. Automatic metros have been used for a long time – being present in over 25 cities. Highly autonomous trains' journeys, on the other hand, started in 2018 in Western Australia, by the Rio Tinto Company, and were a breakthrough. The company claims that by the end of the year, the train has completed more than 1 million km autonomously with remote supervision.

The revolution of autonomous transport modes has reached the maritime sector, as well. Yara Birkeland, an autonomous and electric container vessel developed by Yara and Kongsberg, is expected to go through the first operational tests at the start of 2019, and to conduct fully autonomous operations by 2020

[3] Sources: 2004 DARPA Grand Challenge website (https://archive.darpa.mil/grandchallenge04/index.htm) and https://www.digitaltrends.com/cars/waymo-self-driving-cars-reach-8-million-miles-on-public-roads/

[6]. DNV ReVolt, an unmanned, zero-emission, shortsea concept vessel developed by DNV GL, is being tested in a 1:20 scale, in collaboration with the Norwegian University of Science and Technology (NTNU) [7]. In addition, NTNU is currently testing a 1:2 scaled autonomous passenger ferry, which is expected to run on full scale in 2020 [8].

Prototype of NTNU Autonomous passenger ferry
Photo: Kai T. Dragland / NTNU

In aviation, automation was initially applied in military operations. The Hewitt-Sperry Automatic Airplane first flew in 1917 and was designed as a pilotless aircraft to deliver explosives during World War I. From those early flights, the aviation industry has propelled itself further, with systems such as autopilot and auto-throttle.

Discussion on autonomy in aviation ranges now from autonomous unmanned aerial vehicles (UAV) systems to pilotless commercial aircrafts. Unmanned systems are not only re-shaping transportation systems, but also allowing exploration and research of harsh remote environments with no human life exposure. The Arctic Unmanned Aircraft System Initiative of the Canadian government is testing drones to monitor Canadian Artic for oil spills, ice coverage, marine habitats and activity on the oceans [9]. Unmanned aircraft and remotely operated ground vehicles have also been used to monitor Japan's Fukushima nuclear power plant accident in places too dangerous for humans [10]. Currently, UAVs use range from policing and surveillance to product deliveries and aerial photography. Civilian UAVs now vastly outnumber military UAVs.

Autonomous Underwater Vehicles (AUVs) are also used for tasks in harsh and unstructured environments, such as for ocean monitoring, in detailed mapping of the seafloor, and for inspection of subsea infrastructure. Similarly, autonomous systems have been used in space exploration. NASA has a team responsible for developing a suite of intelligent system technologies to extend ground support for deep-space exploration. In addition, to reduce manpower requirements and account for the time delays in communications, the International Space Station (ISS) incorporates advanced autonomous feature. These include smart sensors for failure recognition, diagnostics and prognostics, model-based reasoning for scheduling maintenance, and automation of low-level routine tasks [11].

The rapid development of the technology-enabling systems with some degree of autonomy is driven by the extensive benefits it brings to the wide range of applications mentioned above. Autonomous systems may bring enhanced solutions to city traffic, cargo transport, data collection and knowledge building of harsh environments, and space exploration. The development of autonomous system applications is, however, not without challenges.

Recent accidents have put emphasis on the need to discuss the safety aspect of these systems. The media has particularly featured recent accidents involving autonomous cars, especially the ones causing fatalities. In 2016, two accidents led to drivers' fatalities, in China and in the United States of America [12,13]. These were followed by two accidents in 2018 in the U.S, which led to a pedestrian fatality and a driver fatality [14,15]. More recently, in January (2018), a self-driving car hit and destroyed a Promobot, an autonomous robot who was attending the Consumer Electronics Show in Las Vegas [16]. The car continued to move for 50 more meters before coming to a halt, leaving the robot non-assisted.

Promobot robot (source: @promobot instagram)

Other incidents involving autonomous systems include, among others, an autonomous bus that collided with a truck in Las Vegas in 2017, an autonomous train that crashed into a wall during a test in India in 2017, a U.S. military drone that was hijacked in 2011. Considering these incidents, development of safe solutions for autonomous systems are, more than ever, crucial for their use. In particular, it is essential to:

- Recognize, understand and assess the risks involved with autonomous systems operations;
- Implement safe solutions in the design phase of these systems;
- Monitor, follow up, and ensure that the risk level is acceptable during operation;
- Establish regulations and procedures that assure safe operations;
- Communicate safety to society in order to establish trust in autonomous systems.

Autonomous systems development: what are the challenges?

A common challenge concerning all autonomous systems refers to safety, reliability and security goals being met. Safety can be defined as the state where freedom from unacceptable risks is achieved, or the condition where a system is successfully operating [17]. Reliability, on the other hand, can be defined as the probability of a system or component working as intended under specified conditions for a specified amount of time [18,19]. It is important to note that reliable systems are not necessarily safe. A reliable autonomous system may execute an action each time perfectly but, in conjunction with external circumstances, such a reliable action can lead to an accident.

The difference between reliability and safety becomes more apparent when the software used in autonomous systems is considered: The software may be executed reliably but may not be safe. For instance, instead of stopping when being operated outside its design envelope, the control software may attempt to recover the system. Similarly, a safe system is not necessarily secure. Security can be defined as the freedom from unacceptable risks being created through voluntarily actions targeting directly or indirectly the system [18,19]. The vulnerabilities that a threat agent exploits arise from within the system or through design flaws. Safety features may be exploited by hostile agents in order to gain control of or access to an autonomous system. Conversely, a secure system may be not safe for users, e.g., due to an over complicated operation.

In the following pages, we will present five key areas that can pose a challenge for SRS of autonomous systems.

Interaction of software, hardware, and human operator

One of the complexities that characterize autonomous systems is the strong interaction among its different components. These are hardware, software, computer hardware and the human operator or supervisor, when applicable. All these interactions occur in a partially unknown and difficult to predict environment. Human operators are often seen as responsible for accidents, either by initiating them or by not responding properly in the course of events. Indeed, one of the motivations for autonomous systems development is their potential to rely less (or not rely at all) on humans for operation and, consequently, for accidents where human failure would be involved to be avoided. However, depending on the LoA of the autonomous systems, it will still rely on humans for remote control, for onboard operation in part of their task, or for monitoring.

In autonomous systems, operators may use system's functionalities out of the intended context or design envelope, or not behave as expected when their actions are required for emergency response. Their interaction with the system may, thus, voluntarily or involuntarily, jeopardize the SRS of the system. Likewise, a failure of the software may provide misleading information to operators or not provide the necessary data, thus leading to human failure. Similarly, the hardware may produce noise or faulty signals that are interpreted incorrectly by the software, which may lead to unanticipated and often unwanted effects. Software, in turn, may not work as intended and lead to faulty activation of actuators or display imprecise information, due to the discrete nature of the software – both in time and enumeration.

Finally, interactions may create vulnerabilities that can be used by malicious agents to take control of the autonomous system. The challenges regarding SRS lie in identifying failures that may arise from this complex interaction, as well as from the propagation of those throughout the system's components and subsystems. Solving this challenge will allow for providing valuable contributions to the identification and development of efficient risk-reducing measures and SRS management strategies.

Assessment methods for safety, reliability and security

The software-hardware-human interaction discussed above is one of the main challenges for SRS assessment of autonomous systems. Most current quantitative assessment methods used in conventional risk and safety assessments rely on the separation principle. System components are assumed to be independent of each other and are often analyzed separately [20]. The interaction among components and emerging complexity is thus often neglected or reduced to a minimum. This makes it possible to use proven methods; however, complex systems may be abstracted and not sufficiently represented.

Some qualitative methods incorporate the different system elements, assessing the emerging properties and system interactions. These are, for example, STPA [21] or FRAM [22]. Such methods, while providing useful qualitative analysis, are still very limited in unravelling complex failure modes and mechanisms in addition to being qualitative and of limited value in prioritizing risks and risk reducing measures. The assessment of hardware with respect to SRS is generally well established. Mathematical approximations of failure probabilities of elements, such as engines, valves, or drive trains, are well developed and publicized. However, computer hardware is subject to different failure mechanisms and patterns and the established methods only apply to a limited extend.

For software, SRS assessments are more difficult to establish. Reliability is approximated by such measures as the remaining amount of errors in the software, which does not clarify how the software may fail. In particular, the interaction of different software components, from possibly different suppliers or development teams, is challenging. Several thousand lines of code need to be analyzed and checked for possible interactions. Risk analysis for software has been addressed recently, which is different from reliability methods [23]. Many of the commonly used approaches for software SRS assessment in the industry build on checklists and or focus on fulfilling formal requirements as proof for SRS compliance [19].

An additional challenge concerns security assessment of AS, including, but not limited to cybersecurity. New threats and vulnerabilities may emerge with autonomous systems. The complexity of autonomous systems may mask vulnerabilities, and attackers may use the complexity to hide their intrusion or access. The assessment of still unencountered threats, malicious intentions and attackers is a key step for addressing security [24].

Autonomous systems are complex, with emerging properties from the interactions of the systems' components. Therefore, a holistic approach is required for the SRS assessment, considering the possible interactions and their potential outcomes and implications [22]. Theory on cyber physical systems and systems of systems may assist in in handling this complexity.

Cyber security

Cyber security, data security, Information Technology (IT) security and physical security may be one of the major challenges concerning autonomous systems. The autonomous behavior may be exploited, and passengers and goods may be endangered. Security addresses the malicious exploitation of vulnerabilities through threat-agents to cause harm or benefit from it. The threat agent may be internal or external to the system. This is often connected with hacking, where software vulnerabilities are abused, and the attacker accesses the target system to control it or extract information [25]. Vulnerabilities are created through the design of hardware or software, the human users, or process related flaws. Hardware hacking is another method to access a computerized system. Microchips or micro computers are introduced in the system and allow an attacker to access the computer system [26].

Practices and components that can create vulnerabilities are shared among different types of autonomous systems, for example, communication protocols between components that have been developed many years ago and do not have any security mechanisms. Vulnerabilities may also arise from poorly-integrated system components, wireless communication and/ or entertainment

systems, interaction of a human user with the system, processes the system is involved in, remote monitoring systems, inadequately trained machine learning systems [24,27–31]. A cyber-attack may not always target the autonomous system itself. A ransom ware or a virus may inflict collateral damage to the autonomous system and disable it.

Although autonomous systems may not have an email address or allow downloading of files, the user or operator may connect to the system using his/her own device. This may open the system for intrusion or give access to malware [32]. Another aspect of cyber security for autonomous systems is jamming and spoofing of sensor systems [24]. A jammed sensor is not able to fulfil its function due to a disturbing signal that disables it. A spoofed sensor, on the other hand, will produce fake signals. Jamming and spoofing may affect, among others, visual sensors, radio wave sensors and global navigation satellite systems. It has been demonstrated that by jamming and spoofing autonomous systems can be hijacked and stolen [25,30].

Autonomous Systems should be developed having in mind these vulnerabilities. A sound cyber security management system is required from early development stages on.

Legal and regulatory aspects

Legal and regulatory aspects may be particularly challenging for unmanned autonomous transport systems. Transport systems are regulated to, above all, assure their safety regarding communities, users and drivers. However, these regulations, when developed, did not contemplate autonomy being introduced in these systems. Regulators are thus facing the challenge of developing or adapting existing regulations to accommodate autonomous and semi-autonomous vehicles (AVs); and to keep up with the pace of technology development. Developers, on the other hand, face the challenge of demonstrating and communicating safety of their systems to regulators.

Autonomous ships are a current example of the abovementioned challenges. Ship operations are broadly regulated by the International Maritime Organization (IMO)[4]. Although having a centralized regulation scheme brings uniformity of regulatory approach, IMO regulations also move slowly. One of the legal issues is the safe manning requirements applicable to merchant vessels. Several conventions require that vessels shall be properly manned to maintain a safe lookout, which is a challenge for unmanned autonomous ships.

[4]IMO develops guidelines, and those are implemented and enforced by each member state.

In general, such requirements may demand major adaptations within current regulations. For instance, the autonomous bus to be adopted in Stavanger, Norway, will have to operate with an employee onboard, in order to comply with Norwegian legislation. This employee must be able to manually override the autonomous controls with a brake button if a dangerous situation occurs.

Road traffic is generally regulated by The Vienna Convention on Road Traffic [33], an international treaty, since 1968. The convention initially stipulated that a human driver must always remain fully in control of and be responsible for the behavior of their vehicle in traffic. The treaty has been signed and ratified by 75 countries, and examples of non-signatory countries include the United States and China. The fact that the U.S. is not a signatory, combined with the possibility of federal states establishing their own legislation, may have influenced that it was one of the pioneers in legislation for autonomous cars. Nevada was the first US state to authorize the operation of autonomous vehicles, in 2011. Since then, 21 other states have passed legislation related to autonomous vehicles. Recently, the US National Highway and Transportation Safety Administration (NHTSA) released new federal guidelines for automated driving systems (ADS). It has a voluntary nature, without compliance requirement or enforcement mechanism.

In December 2016, an act implementing an amendment to the Vienna Convention on Road Traffic entered into force in Germany [34]. The amendment allows the transfer of driving tasks to the vehicle itself, provided that the technologies used are in conformity with the United Nations vehicle regulations or can be overridden or switched off by the driver. Once again, a licensed driver is required to be behind the wheel to take control if necessary.

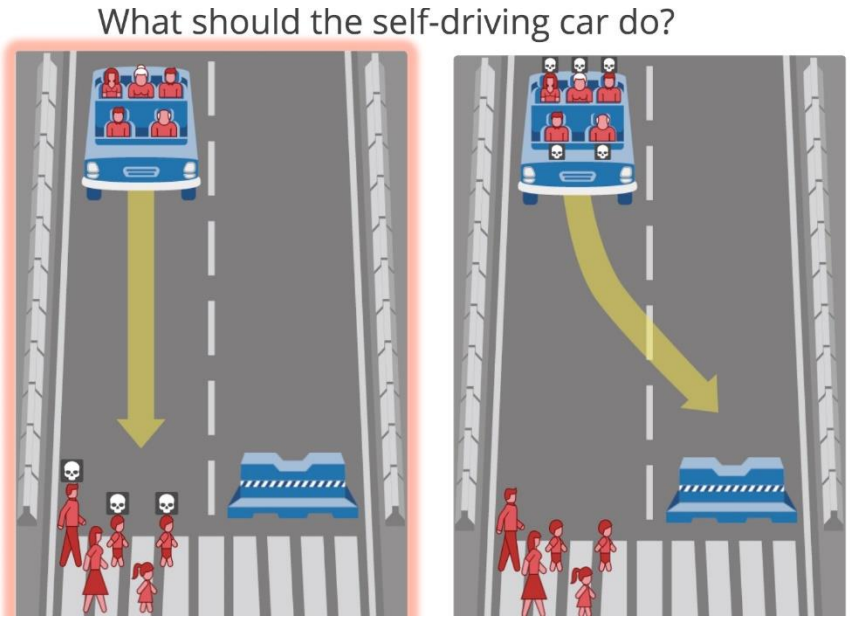
Liability is another challenge in regulating AV. Who should be responsible when an accident happens? Will anti-collision algorithms developers be responsible when a collision occurs? To what extent is the remote driver or supervisor responsible in case s/he does not act in time to override an action from a mal-functioning system?

In addition to the questions above, some ethical aspects must be assessed in terms of liability. For instance, in the U.S., the income of the victim is related to her/his liability damages – the more someone earns, the greater her/his liability exposure. To protect themselves against major liability claims, AV manufacturers may adjust the car's driving behavior according to the average income in an area [35]. The problem of regulations for autonomous vehicles comes with a catch-22: we need to test and use AVs to assess their safety; yet we do not want them on the road / ocean / sky until we know that they are safe.

Ethical and social aspects

"Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation." [36]

The above quote is retrieved from the report of the developers of the Moral Machine[5]. The experiment, launched in a website, was developed to collect large-scale data on how people would want autonomous vehicles to solve moral dilemmas. The interest in the platform was significant, and they collected almost 40 million decisions from nearly all countries of the world. The experiment presents users with an unavoidable accident scenario and offers them the choice of the car to swerve or stay in course. The outcome of this choice is to spare one group over the other during a collision; for instance, if the car stays in course it may run over pedestrians, and if swerving it will collide with a fixed object and danger the passengers. They collected decisions data over nine main factors, as sparing men versus women, or humans versus pets.



Moral Machine (source: http://moralmachine.mit.edu/)

The type of choice the users confronted in the Moral Machine follows the framework of the trolley cases and has been addressed by ethics researchers on analyzing autonomous cars. The choice on who to harm in case of unavoidable accidents is a necessary question regarding the development of autonomous vehicles. Should this decision be fixed and embedded in the algorithms during

[5] http://moralmachine.mit.edu/

development? Will cars use machine learning and "replicate" human-alike decisions? These questions become more difficult to address given the results of the Moral Machine experiment. Although there were some consensuses regarding some dilemmas, as sparing humans over animals; significant socio-geographical differences arose when dealing with other choices. For instance, a preference to spare younger characters/ people is less pronounced in far eastern countries and in some Islamic countries, and higher in Latin America. The same is true for the preference in sparing higher status characters [36].

Imitating human drivers' behavior for establishing moral decisions is, thus, a challenge given the socio-geographical differences. In addition, humans may show unethical biases when driving, such as deciding whether to yield at crosswalks based on pedestrians' race and income [37]. Ethics of autonomous vehicles are not restricted to the trolley problem [38]. Mundane traffic situations, such as approaching a crosswalk with limited visibility, making a turn, navigating through busy intersections, or factors related to how liability is determined raise important ethical question [35].

The first and only attempt so far to provide official guidelines for the ethical choices of autonomous vehicles is the German Ethics Commission on Automated and Connected Driving [39]. One of the rules states that, in a dilemma, protection of human life should have priority over other animals' life. Another rule affirms that distinction based on personal features such as age, should be prohibited. How ethics and moral are implemented on AS will influence its societal acceptance. People's willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted. In addition to moral aspects, trust in autonomy is an important factor for societal acceptance. Trust in automation is a highly discussed subject in the human factors and human reliability community.

In short, autonomy creates a new depth in the human-machine relationship from the users' side, the operators that supervise it or remotely control it, and the people interacting with the autonomous systems externally. Communicating safety to society is thus a must to gain trust in autonomy and societal acceptance.

References

[1] Vagia M, Transeth AA, Fjerdingen SA. A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? Appl Ergon 2016;53:190–202. doi:10.1016/j.apergo.2015.09.013.

[2] Clough BT, Leader TA, Force A, Afb W. FAO backs biotech crops. Inf - Int News Fats, Oils Relat Mater 2004;15:438.

[3] Sheridan TB, Verplank W. Human and Computer Control of Undersea Teleoperators. Cambridge: 1978.

[4] Sheridan TB. Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: Distinctions and modes of adaptation. IEEE Trans Syst Man, Cybern Part ASystems Humans 2011. doi:10.1109/TSMCA.2010.2093888.

[5] Laurinen M. Remote and Autonomous Ships: The next steps. AAWA Adv Auton Waterborne Appl 2016:88.

[6] Kongsberg. Autonomous ship project, key facts about YARA Birkeland 2017. https://www.km.kongsberg.com/ks/web/nokbg0240.nsf/AllWeb/4B8113B707A50A4FC125811D00407045?OpenDocument.

[7] Alfheim H, Muggerud K. Development of a Dynamic Positioning System for the ReVolt Model Ship Henrik Alfheim. NTNU, 2017.

[8] NTNU. Autoferry – Autonomous all-electric passenger ferries for urban water transport 2018. https://www.ntnu.edu/autoferry (accessed August 13, 2018).

[9] Transport Canada. Drones in the Canadian Arctic n.d. https://www.tc.gc.ca/en/programs-policies/programs/national-aerial-surveillance-program/drones-canadian-arctic.html (accessed February 1, 2019).

[10] Lillian B. Engineers Developing Drones to Inspect Fukushima Daiichi Nuclear Disaster n.d. https://unmanned-aerial.com/engineers-developing-drones-to-inspect-fukushima-daiichi-nuclear-disaster (accessed February 1, 2019).

[11] Committee for NASA Technology Roadmaps Aeronautics and Space Engineering Board Division on Engineering and Physical Sciences. Autonomy research for civil aviation torward a new era of flight. Washin: 2014.

[12] Boudette NE. Autopilot Cited in Death of Chinese Tesla Driver. New York Times 2016.

[13] Singhvi A, Russell K. Inside the Self-Driving Tesla Fatal Accident. New York Times 2016.

[14] T.S. Why Uber's self-driving car killed a pedestrian. Econ 2018.

[15] Ohnsman A. Fatal Tesla Crash Exposes Gap In Automaker's Use Of Car Data. Forbes 2018.

[16] Cozzens T. Autonomous car hits autonomous robot in bizarre collision n.d. https://www.gpsworld.com/autonomous-car-hits-autonomous-robot-in-bizarre-collision/ (accessed February 1, 2019).

[17] Woods DD. Essential characteristics of Resilience. In: Hollnagel E, Woods DD, Leveseon NG, editors. Resil. Eng. -Concepts Precepts. 1st ed., Surrey, UK;

Burlington, USA: Ashgate; 2006, p. 21–34.

[18] Rausand M. Risk Assessment - Theory, Methods, and Applications. 1st Ed. Hoboken, New Jersey, USA: John Wiley & Sons; 2011.

[19] ISO, IEC. ISO/IEC Guide 51: Safety Aspects - Guidelines for their inclusion in standards. Geneva, Switzerland: 2014.

[20] Mosleh A. PRA: A Perspective on strengths, current Limitations, and possible improvements. Nucl Eng Technol 2014. doi:10.5516/NET.03.2014.700.

[21] Leveson NG, Thomas JP. STPA Handbook. 1. Cambridge, MA, USA: 2018.

[22] Hollnagel E. FRAM – The Functional Resonance Analysis Method. 1st Ed. Farnham. UK: Ashgate; 2012.

[23] Aldemir T, Guarro S, Mandelli D, Kirschenbaum J, Mangan LA, Bucci P, et al. Probabilistic risk assessment modeling of digital instrumentation and control systems using two dynamic methodologies. Reliab Eng Syst Saf 2010. doi:10.1016/j.ress.2010.04.011.

[24] Petit J, Shladover SE. Potential Cyberattacks on Automated Vehicles. IEEE Trans Intell Transp Syst 2015. doi:10.1109/TITS.2014.2342271.

[25] Yağdereli E, Gemci C, Aktaş AZ. A study on cyber-security of autonomous and unmanned vehicles. J Def Model Simul 2015. doi:10.1177/1548512915575803.

[26] Wyglinski AM, Huang X, Padir T, Lai L, Eisenbarth TR, Venkatasubramanian K. Security of autonomous systems employing embedded computing and sensors. IEEE Micro 2013. doi:10.1109/MM.2013.18.

[27] Bothur D, Zheng G, Valli C. A critical analysis of security vulnerabilities and countermeasures in a smart ship system. 2017.

[28] Haas RE, Möller DPF. Automotive connectivity, cyber attack scenarios and automotive cyber security. IEEE Int. Conf. Electro Inf. Technol., 2017. doi:10.1109/EIT.2017.8053441.

[29] Hassani NTNU V, Ocean S, Trondheim ntnuno, AntónioAnt P, Pascoal AM. CYBER SECURITY ISSUES IN NAVIGATION SYSTEMS OF MARINE VESSELS FROM A CONTROL PERSPECTIVE. 2017.

[30] Parkinson S, Ward P, Wilson K, Miller J. Cyber Threats Facing Autonomous and Connected Vehicles: Future Challenges. IEEE Trans Intell Transp Syst 2017. doi:10.1109/TITS.2017.2665968.

[31] Vinnem JE, Utne IB. Risk from cyberattacks on autonomous ships. Saf. Reliab. – Safe Soc. a Chang. World, 2018. doi:10.1201/9781351174664-188.

[32] Cárdenas AA, Amin S, Sinopoli B, Giani A, Perrig A, Sastry S. Challenges for Securing Cyber Physical Systems. n.d.

[33] UNITED NATIONS CONFERENCE ON ROAD TRAFFIC. Convention on Road Traffic. Vienna: 1968.

[34] Amendments to Article 8 and Article 39 of 1968 Convention on Road Traffic 2016;2016:1306–8.

[35] Himmelreich J. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. Ethical Theory Moral Pract 2018;21:669–84.

doi:10.1007/s10677-018-9896-4.

[36] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. Nature 2018;563:59–64. doi:10.1038/s41586-018-0637-6.

[37] Coughenour C, Clark S, Singh A, Claw E, Abelar J, Huebner J. Examining racial bias as a potential factor in pedestrian crashes. Accid Anal Prev 2017;98:96–100. doi:10.1016/j.aap.2016.09.031.

[38] Roff H. The folly of trolleys: Ethical challenges and autonomous vehicles. Brookings 2018.

[39] Federal Ministry of Transport and Digital Infrastructure of Germany. Ethics Commission on Automated and Connected Driving. 2017. doi:10.1126/science.186.4158.38.

Autonomous Transportation Technologies: Society and Individuals in the Loop

Report of the Discussions of Breakout Session

Authors: Daniel Metlay (session chair), Asun Lera St. Clair, Bernhard Twomey, Jens Einar Bremnes, Ørnulf Rødseth, Salvatore Massaiu, Siri Granum Carson, Stig Ole Johnsen, Thomas Porathe

Social Acceptance and Trust

Autonomous transportation systems strive to replace human decision-making with computer decision-making. For the immediate future, this transformation takes place in an open system in which human decision-makers are likely to be present.

Out of necessity, autonomous technologies, like all other technologies, will contain embedded values. Engineers, scientists, designers, regulators, sponsors all make choices that, either intentionally or unintentionally, enhance certain cultural and societal values [1]. Consequently, "pure" or "morally neutral" technologies are rarely, if ever, deployed. The more pervasive the technological system, the more profoundly will be the "techno-moral changes" it stimulates [2].

Widespread implementation of virtually all technologies is likely to differentially affect how values are distributed. Sufficient attention to and self-conscious reflection on this inevitable dynamic is necessary to secure socially responsible research and innovation in the area of autonomous technology. Only in that way will it become transparent what specific values are subverted, supported, or unaffected. In particular, recent developments in the area of autonomous systems technology, such as deep learning, come with certain challenges. While transparency and accountability are key values for any approach to technological development, these values may be harder to uphold for systems where the possibilities for *explicability* of the involved operations are limited. Thus, Floridi et al. [3] suggest that for these forms of technology, explicability in the sense of a need to "*understand* and *hold to account* the decision-making processes" is a key value.

Ultimately, social acceptance of any new technology depends on trust, both in the technology itself and in the organizations that implement and regulate it. Trust in a technology requires that it not produce significant unpleasant surprises. Trust in an organization requires that it demonstrates both

competence and fairness. Put differently, a trustworthy organization knows what it is doing and acts in a way that is consistent with the values of the large society.

Social acceptance of autonomous transportation technologies requires that open and inclusive processes be put in place in order to identify and evaluate what values are embedded in the technology and what values are affected by it. The framework developed by Stilgoe et al [4] describes for dimensions of responsible innovation: anticipation, reflexivity, inclusion and responsiveness. These are intended as general rules of thumb to guide the governance of socially responsive and responsible innovation, and this framework has become influential in European research and innovation policy.

In the next section, the question of governance is explored in greater depth.

Governance

Governance refers to all acts of managing technology, whether by laws and regulations from governments of a country or groups of countries or by the market, industry standards, or by a network or group (such as a tribe or family). The term also implies that there is a degree of self-regulation by societal actors and to private-public cooperation in solving societal problems. Because governance is concerned with realizing public goals, it always entails a process of steering (regulating) a particular constituency of actors and is regarded as authoritative and legitimate.

Autonomous systems are receiving close scrutiny because they hold the potential to incur high societal impact (both positive and negative) and because of the ethical issues posed by the substitution of human decision-making for machine decision-making. In particular, artificial intelligence (AI) and machine-learning components of autonomous systems are under the scrutiny of public regulators and society at large. This attention is warranted because safety is harder to assure, as these technologies exhibit "emergent" behaviors that are *a priori* unpredictable and can lead to unanticipated failures.

Governance can be public, private, or a hybrid of both. Public governance is most commonly exercised by public institutions, government agencies, at different scales (local, national, regional, international or global). Public governance is often legally binding and entails liability. Public governance regulates the behaviors or people and organizations, by setting procedures and constrains for products. Public governance is a reflection of the social contract between citizens and those governing, reflecting the values and societal expectations of citizens. Presumably, individuals and organizations possessing

skills related to autonomous systems such as design and testing will be called upon more frequently to contribute their knowledge to effective governance of these systems.

Private governance emerges when private actors self-organize and create rules and recommendations. These are voluntary and often emerge from a combination of self-interest, cooperation and negotiation. Private governance operates via industry rules and regulations, best practices, and standards. Often private governance is simply the result of a public organization delegating a governing role to a private actor. There may also be circumstances where individuals and organizations possessing skills related to autonomous systems act independently or in unison to ensure desirable properties of these systems are achieved or negative outcomes prevented. Examples include consortia of companies, groups of concerned scientists, and similar groups of experts who do not hold equal access to the reigns of regulatory power.

Very often, public and private governance meld to form hybrid governance, with elements of both public and private actors, resulting from their interactions, negotiations, and collaborations. An example of a hybrid governance system is the maritime safety regime which has elements for national (flag states), global (International Maritime Organization [IMO]), and private actors (shipping industry or Class societies). (The latter is a hybrid organization, to whom governmental organizations delegate power and authority to carry on governance roles, but that are also dependent from and interacting with the maritime industry players.)

The governance of autonomous systems must encompass all these different forms of governance. One governance goal may be to map a path toward more nimble forms of autonomous, potentially partially or fully autonomous systems that preserve desirable attributes such as safety while increasing the potential for utility through deployment of services valued by consumers and citizens.

Historically, it could be argued that regulatory power has been consolidated in the hands of a minority. Pluralistic systems of government have experienced convergence to equilibria in which corporate interest disproportionately influence regulation through monetary control. In these environments, business interests take precedence over stakeholders who are not represented fairly nor protected in an equitable manner. Mechanisms to disassemble these structural inequalities in an orderly manner will enable preservation of social stability. However, this transition will likely experience uneven progress and become the focal point of fierce debate.

Governance challenges

Autonomous transportation technologies present very serious challenges to traditional forms of governance due to:

- their being in an early stage of development but in need of testing;
- the independent decision making and the changing nature of these systems;
- the technological competence needed by regulators;
- the fast pace of technological advance.

Most laws address human behaviors and are typically inadequate to deal with complex human-machine interactions. Machine-machine and human-machine interactions will (may) provide value lessons on how humans should treat each other and behave.

Autonomous systems require adaptable and agile governance regimes. These need to be iterative and ongoing through the whole lifecycle of systems and cover all the human- machine-interaction elements. Software updates can change the nature and capabilities of a system; even if the systems had met the criteria of regulators and the country where it operated before the software update, it may no longer do so after the update is implemented.

These interactions would benefit from formal modelling. This strategy would enable incremental approval for progressively more widespread use of systems in environments where they come into contact with humans. It would also help constrain how these interactions take place to decrease the likelihood of safety related incidents and other undesirable outcomes.

At least part of the solution could be achieved through software frameworks that explicitly codify discrepancies at various levels of jurisdiction such as transnational, national, or state/province. These software frameworks may need to be sufficiently complete to ensure continuous operation of systems to avoid creation of safety and other compliance related issues. For example, to ensure laws are observed, autonomous vehicles approaching a national border may wake the human to take control because the country they are entering only permits a lower level of autonomous driving. An autonomous vehicle could rely on existing technologies such as geofencing to verify safety and other attributes through test cases. Methods that attempt to enumerate combinations of scenarios could provide a process by which gaps in these rules are filled with explicit guidance. The severity of these gaps could also be specified to prioritize consideration of their implications as well as to search for other areas where such scenarios may arise.

Simulation technologies (including human-machine interactions) would also inform such testing to expose gaps in regulatory policy. These efforts could be informed by past research in expert systems (now referred to as first generation or rules-based artificial intelligence).

Technological development ahead of governance

Technological development outpaces current governance systems for many industries and sectors. This circumstance may continue to be the case for the foreseeable future until alternative mechanisms that are sufficiently robust can be designed and implemented. Moreover, there are major gaps in the regulation for autonomous technologies. Existing regulations prevent the development or a license to operate of autonomous systems. These restrictions will likely be contested by organizations that stand to gain financially or otherwise from the deployment or marketing of these systems.

Autonomous systems also pose challenges to the current governance of many sectors. For example, although autonomy requires collaboration at the international level, organizations like the IMO may not have sole jurisdiction in autonomous systems. Thus, while the IMO can regulate shipping at sea, autonomous shipping will likely have elements of their system (control operators) on land in a particular country. Accordingly, the IMO and every nation state that has land operations would have to agree on a common set of regulations. The ecosystem of governance participants will likely advocate for reforms that allow them to protect their own interests, including economic and security concerns.

At the same time, it is important to acknowledge that there are existing rules and regulations that are applicable or can be easily revised to address autonomy issues. Understanding which ones those are and identifying critical gaps may be a challenge. Formal methods could ensure this is accomplished in a repeatable and correct manner.

The processes of creating rules and regulations (whether public laws or the development of new standards) for autonomous systems needs to be inclusive and democratic as well as reflect the diversity of concerns of different stakeholders. Mechanisms to hear and respond to the voice of under-represented stakeholders are not well developed. Procedures to enhance opportunities for input from those individuals could provide broadly beneficial. Without avenues to collect missing inputs to inform decision-makers, corporate interests will continue to exert outsized influence on regulation to ensure future profits, but this circumstance compromises the integrity of democratic value elicitation to guide resource allocation.

Due to the increased complexity of autonomous systems, moves are already underway to develop goal-based rules and regulations (emerging from hybrid governance). More research is needed, in collaboration with the industry, to identify a new generation of assurance methods for complex and intelligent systems. These methods may be informed by social media and technologies such as natural language processing. Crowdsource sentiment analysis can provide more fine-grained details to characterize stakeholder values.

Automation poses key questions for third party assurance providers, as the cyber component of systems is often not part of the mandatory requirements or often overlooked or incomplete. Certification organizations within a particular industry sector need to quickly recruit additional expertise to ensure they have both digital and domain competence to deal with new digital risks. In some cases, it will not be possible to market or deploy products without sufficient understanding of the system to be able to attribute liability. This may in turn require that regulators restrict the level of autonomy and emergent behavior that a system exhibits until developers can demonstrate the technologies have achieved an acceptable level of maturity and that risks have been mitigated in a reasonable manner. Spaceflight and the nuclear power industries are two examples where heavy regulations are enforced to ensure human safety and environmental protection.

A key aspect in the governance of autonomous systems relates to the training of needed skills required to design, operate, maintain and evaluate autonomous or semi-autonomous systems. The accidents of the Boeing 737 Max offer glaring examples of this point.

Automation can lead to changes in business models, and this shift means that new and existing actors take on different roles. For example, a ship owner or operator may become a transport company. New rules and regulations will have to address this possibility. Autonomy is a disruptive technology and, like any industry, it is important to understand whom the players in a particular autonomous system are and how responsibilities need to be shared.

Ethical and societal challenges

Because autonomous systems are likely to introduce ethical and societal challenges, it is essential to pay attention to the societal impacts in the emerging governance of autonomous systems.[6]

[6] See for example the US NSF's 10 Big Ideas (https://www.nsf.gov/news/special_reports/big_ideas/) includes the future of work at the human-technology frontier.

Autonomous systems hold promise to enhance society broadly including quality of life, environment, and simply greater convenience. However, the traditional extended process of defining regulatory procedures may be poorly stressed as corporate entities developing autonomous systems aggressively challenge how long it takes to market a new technology.

This conflict should raise citizen pressure to ensure that all stakeholders, underrepresented or not, are properly engaged in all domains in which autonomous systems impact their lives. Without such engagement, issues that have been raised in the context of racially biased machine learning algorithms are likely to repeat themselves. Other instances noted in the media include gender discrimination in the display of job advertisements.

It has been presumed that autonomous systems will always be used for good purposes and will hold the promise of enhanced quality of life. However, more oppressive uses of such technologies seek to assert social control through surveillance and monitoring to enforce conformance. Most commentary assumes that no party will seek to subjugate the majority of humans to machines. More positive examples might include a holistic educational system that retrains individuals based on their educational profile after automation renders their present occupation obsolete. Past research in intelligent tutoring systems could inform this process and be coupled with more recent technologies such as virtual and augmented reality.

How humans regulate autonomous systems technology can serve as a model to understand how humans regulate their behavior and the discrepancies involved in the inequitable application of laws. Not only would humans like to act and autonomous systems to follow, they would also like regulation/ laws to follow according to collective valuations tempered by the trade-off between utility and corresponding risks. For example, autonomous vehicles are typical trained on images representative of developers, predominantly fair skinned men. As a result, autonomous vision systems have lower detection rates of other populations such as women wearing skirts and dark skinned people. Thus, while autonomous systems could reduce driver related fatalities, some populations may be disproportionally impacted by the bias in the training data. Hence, if the deficiencies in reliability and safety of technologies go undetected, we may experience rapid and unanticipated shifts that results in a substantial increase or decrease in stakeholder acceptance of policies imposed.

We need regulation that enables testing but at the same time enhances learning and safety. However, not all systems enable completely safe learning. Some types of systems are designed to take calculated risks to learn from their exploratory actions. Too much regulation at the early stages of technological

maturity may have negative impacts on the ability of these systems to mature in their intended environments. At the same time, a rush to deploy products in the market and the prioritization of industry interests by regulators may lead to accidents or failures with unacceptable societal and ethical consequences.

Autonomous Transparency

Automation transparency describes how the automated system communicates with humans to ensure mutual understanding and promote good "teamwork." How the system presents information to individuals in the environment will be critical for the acceptance and safety of the system. Autonomous systems will constantly interact with a variety of individuals:

- Humans in the environment outside the autonomous vehicle,
- Humans inside the system if it for example transport passengers or
- If the system is not fully autonomous, remote operators in a control center.

An example from the marine domain is the small passenger ferry that is currently being developed by the Norwegian University of Science and Technology in Trondheim. The ferry will go back and forward across a 100 meters-wide harbor canal. The passage time will be around one minute. This task might seem simple at first, but it contains several complicating issues relating to the three bullet points noted above:

The canal is not a controlled space like an elevator shaft. Instead there are both commercial and leisure traffic, especially in the summer time. Even if the ferry is programmed to follow the international collision regulations ("COLREGS") we cannot rely on tourists in the kayaks and every boater are familiar with these rules. The ferry needs to communicate its intentions in an unambiguous way. It also needs to behave in a way that is understandable and makes sense to humans in other boats – even if such behavior is not optimal from an efficiency point of view.

The ferry automation also needs to communicate with the passengers onboard. The communication extends from counting passengers embarking and disembarking to informing passengers about the regular safety issues and regular performance. But the most important and difficult task will be to handle emergency situations without crew onboard, such as if the ferry needs to be evacuated on the canal.

The ferry is monitored from a remote location. The human-machine interaction between the automation and the human remote operator will be crucial. From the start, this monitoring will be very watchful, but as systems are expected to become more and more reliable we expect the monitoring to become

more relaxed and eventually we envision that the remote monitoring to be switched over to a more limited environment and also a small tablet based mobile control environment.

Human-Machine Interactions

Humans have, when faced with completely new situations, the ability to adapt to the new situations and improvise to handle the situation appropriately. These skills are especially important for high-consequence scenarios. Current state-of-the-art autonomous systems however, have limited adaptation and creativity skills and will in the future be required to improve them.

Does machine learning have what it takes when it comes to high-risk systems in terms of adaptation and creativity? Machine learning algorithms build a mathematical model of sample data, often referred to as training data, in order to make decisions or predictions. In order to properly train a machine-learning model, large quantities of data are needed.

For low-probability scenarios, limited or no data are available. In such cases, the uncertainty associated with the predictions will be substantial. This situation will significantly reduce the predictive accuracy of the machine-learning model for high-consequence scenarios. In those cases, the tolerance for erroneous predictions should be low, as faulty predictions may have severe consequences.

Machine learning algorithms aim to find patterns and causalities directly from historical data. Therefore, machine learning models will, by definition, be held tightly to previous events. Relying on purely data-driven models for safety-critical systems may have important limitations.

Adaptation and creativity may involve simulating and evaluating all possible actions and choosing the best feasible action. However, as the state-space increases with the complexity of the systems, evaluating all possible actions may will not be feasible in real-time. In order to enable real-time simulation-based decision-making, it may be necessary to abstract the system, using heuristics to guide the simulations and state-space searches, or use simplified surrogate models.

The Role of the Operator in Autonomous Transportation Technologies

The primary reason for including human operators in autonomous systems is to cope with situations that are beyond the reach of automation or designers' foresight [5]. An effective autonomous system means that previously considered "beyond the reach situations" are now considered within the control boundaries of the automation; ideal (or acceptable) system performance is

assumed also in environments (and for users) formerly considered unpredictable or intractable.

Yet, for all activities that cannot be completely reduced to a set of algorithms (e.g., transients, extremes and rare cases) human operators (and even users) may still be required to perform two functions: (1) discretionary decision-making and (2) recovery, i.e., response in case of failures, including automation failures. This means that highly automated systems' productivity and safety will be dependent on their capability to support the necessary human intellectual abilities underlying the two 'adaptive' functions. Operator support in this sense is beyond good ergonomics (usability concerns) and also beyond the models and analysis techniques developed in control engineering. The expertise of researchers from a wide variety of disciplines including human factors engineering, industrial/ organizational psychology, management science, sociology, anthropology and information science, is called upon.

Highly automated systems modify the very role of the operator from one of active controller and effector to that of a passive supervisor whose primary responsibility is to monitor the system and assume manual control if the automation should fail. In this passive role, operators might experience problems with understanding what the automation is doing (automation induced surprises) and knowing when to intervene [6]. The term "out-of-the-loop" (OOTL) has been coined to refer to such performance problems [7,8].

Automation failures, when they occur, will be very difficult to deal with. Failure detection is an inherently ambiguous situation, one in which there are no straightforward ordering of priorities between (1) actions to ensure safety, (2) actions to maintain production, and (3) diagnostic actions to identify the causes and location of the fault. In the first place, failures might not be detected due to complacency (over-trust): the more reliable the automation is, the more it is trusted and the higher the operators' expectations that the automation will not fail. Reliable automation will also be less monitored and the operators less vigilant, in which case it will be harder to intervene when a failure is detected due to reduced situation awareness.

Independently of complacency and vigilance, automation leaves the operators less aware of the control actions made, as it is easier to remember own actions than if witnessed on other agents. Furthermore, the operators might no longer have the skills called upon in a recovery if the automation has effectively replaced the worker and this is not trained recurrently on manual intervention.

Introducing autonomous systems requires a thorough analysis of the tasks previously allocated to workers as well as the ones that will continue to be

allocated. This is not a trivial evaluation. When new types of automation are introduced in complex human-machine systems the operators' roles and tasks change, often in unanticipated and negative ways [9,10]. The analysis requires considering all human and organizational factors encompassed by the system, identifying the successful operators' heuristics and adaptations, as well as uncovering the informal skills of the trade formerly taken for granted (important operator tasks seldom described in formal documents).

References

[1] Rip A, Schot J, Misa TJ, editors. Managing technology in society: the approach of constructive technology assessment. New York: Pinter Publishers; 1995.

[2] Swierstra T. Nanotechnology and technomoral change. Etica e Polit 2013;15:200–19.

[3] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach 2018;28:689–707. doi:10.1007/s11023-018-9482-5.

[4] Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. Res Policy 2013;42:1568–80. doi:10.1016/j.respol.2013.05.008.

[5] Vicente KJ. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. CRC Press; 1999.

[6] Sarter N., Woods D, Billings CE. Automation Surprises. In: Salvendy G, editor. Handb. Hum. Factors Ergon., John Wiley & Sons; 1997.

[7] Endsley MR, Kiris EO. The Out-of-the-Loop Performance Problem and Level of Control in Automation. Hum Factors J Hum Factors Ergon Soc 1995;37:381–94. doi:10.1518/001872095779064555.

[8] Endsley MR, Kaber DB. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 1999. doi:10.1080/001401399185595.

[9] Dekker SWA, Woods DD. MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. Cogn Technol Work 2002;4:240–4. doi:10.1007/s101110200022.

[10] Woods DD. Decomposing Automation: Apparent Simplicity, Real Complexity. In: Parasuraman R, Mouloua M, editors. Autom. Technol. Hum. Perform., CRC Press; 1996.

Group Participants

Asun Lera St. Clair
DNV GL, Norway

Bernhard Twomey
Rolls Royce Marine, Norway

Daniel Metlay
B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA

Jens Einar Bremnes
Department of Marine Technology, NTNU, Norway

Lance Fiondella
University of Massachusetts, USA

Ørnulf Rødseth
SINTEF Ocean, Norway

Salvatore Massaiu
Institute for Energy Technology, Norway

Siri Granum Carson
Department of Philosophy and Religious Studies, NTNU, Norway

Stig Ole Johnsen
SINTEF Digital, Norway

Thomas Porathe
Department of Design, NTNU, Norway

Safety, Reliability and Security Modelling and Methods for Autonomous Systems

Report of the Discussions of Breakout Session

Authors: John Andrews (Session chair), Stein Haugen, Marilia A. Ramos, Christoph A. Thieme

Assessment of Safety, Reliability and Security of autonomous systems

The introduction of autonomy into systems adds new layers of complexity regarding safety, reliability and security (SRS). The main challenges related to methods for SRS assessment concern the following: i) the adequacy of existing methods; ii) the integrated modelling of hardware, human, software and human interaction; iii) self-learning systems; and iv) data requirements.

Risk Assessment

The general goal of risk assessment is to identify hazardous events, prevent their occurrence and mitigate their consequences. A broadly accepted definition of risk is the expected likelihood of a hazardous event combined with the expected consequences. Important questions arise regarding this definition, such as if risk (in the sense of statistically expected loss) is a relevant measure. Further, is risk related to autonomous systems the same as for traditional systems?

The assessment of the likelihood or frequency of events involving autonomous systems is the most challenging part of risk analysis. An additional challenge concerns risk related to software aspects. Methods to investigate hazards resulting from hardware failure and human error are relatively mature; however, the same is not true for software implementation. The assessment must address not only the occurrences of incorrect responses from the software, but also the failure mode that this induces in the system. Yet, most software reliability methodologies focus on the number of bugs remaining in a code, regardless of their effect on the system. In addition, unlike hardware, the historical performance of a software cannot be considered as indicative of future performance. For autonomous systems the problem is compounded by the fact the software can incorporate self-learning and there is no clear rule-based algorithm to examine. This last property may make a systematic evaluation of the potential hazards problematic and hence the risk quantification breaks down.

Regarding consequences, they may be mostly the same as for non-autonomous systems. Exceptions are systems that currently operate manned and that may become unmanned with the introduction of autonomy, e.g., offshore platforms. In this case, the consequence of an accident could potentially be reduced, given that no life of workers would be threatened. Nevertheless, the negative environmental impact would remain the same.

Risk assessment may be adapted for different applications. Traditional hazards, such as fire and collision, are present in existing frameworks and should be included in autonomous systems' risk assessment. However, autonomy includes new hazardous events for which the risk community must investigate the possibility to address and incorporate in the existing frameworks. Threats due to system connectivity, such as cybersecurity, may be challenging to incorporate in traditional risk assessment frameworks. In particular, the frequency of such events may be difficult to define. Hence, a cooperation and exchange of methods and approaches between industries and application areas is highly necessary.

Reliability and availability

For safety critical applications, there is commonly a strategy for component failures to be 'fail safe'. Whilst enhancing safety, this can have a detrimental effect on reliability. Autonomous systems must be reliable over the time of their mission, as there may not be any option for repair during the mission.

Resilience is indicative of how the system can bounce back after a problem has occurred and therefore may also provide a useful measure of system performance. Availability also indicates the ability of a system to return quickly to the working state following a failure. In the maritime sector, it is advantageous to have a high likelihood of completing several missions for the vessel to be returned to a dock with the required maintenance facilities to prepare the vessel for its next sequence of missions. This is a similar concept to the Maintenance Free Operating Periods proposed by the aeronautical industry.

In addition to being resilient, it is beneficial if autonomous systems can include some form of self-repair. The analysis should not solely rely on probabilities or frequencies, since there is a lack of data and a need for make assumptions in all cases.

Security

Security is related to threats from external agents who have the intention to harm the system. Attacks on autonomous systems can exploit some weaknesses that are particular to those systems, and difficult to foreseen. The

assessment of resilience strongly correlates with security: can the system operate after a security breach?

In security, one needs to look at the different realms, including human, software, hardware, society. One of the most frequently used attack methods is to "hack" the human since this may be the weakest link. Humans can also be a security barrier, and their effectiveness needs to be assessed.

In addition to using humans as a breach for an attack, a concern regarding security of autonomous systems are cyberattacks. During a cyberattack, hackers first scan the system and find an open "port" or a vulnerability. They attempt to get the credentials to infiltrate the system. A challenge concerning this type of attack is that while the hacker needs only one port in, the defender must defend all ports. It is therefore necessary to identify which ports are insufficiently protected. A probabilistic method may help in this identification. Also, the analysis of security can leverage from other application areas.

One approach to security analysis is to use game theory. Other methods are attack trees, expert judgements and scenario roleplays. Vulnerability analysis should be a part of security analysis. This is an essential part of security risk assessment in several industries.

Adequacy of current modelling approaches

One of the key topics in the discussion of SRS assessment for autonomous systems is the adequacy of the existing modelling approaches. Could existing approaches be applied directly? The difficulty in obtaining frequencies for some of the events would be an issue and so existing modelling frameworks can work for part of the system assessment only. However, since these methods have served well in the past, and are relatively efficient for existing systems, they should not be dismissed for autonomous systems. Rather, they can be enhanced with additional assessments for the command and control structures.

The types of accidents that can occur in autonomous systems may not be different from conventional systems; yet, the causes to the accidents will change. Autonomous ships, for instance, differ from a traditional ship mainly regarding the responsible agent for decision: with autonomy, some or all the decision-making processes are moved from a human to software. Also, system design and maintenance for autonomous systems may not be direct "copies" of their non-autonomous predecessor (e.g., additional redundancy, predictive or preventive maintenance may be needed to ensure adequate mission reliability).

Qualitative assessment methods are believed to be largely applicable, although improved methods to move further "to the left in the bowtie" are

required, with a focus on the causal analysis. Security and "software failures", however, pose significant challenges, as stated in the previous section. The challenges are essentially related to two aspects: (i) failure to identify all the circumstances that the software need to be able to handle, and (ii) failure to understand how the software works in all circumstances. While the first challenge is related to hazard identification methods, the second is closely related to the self-learning aspects of the software, which provide challenges in verification and testing.

The identification of all the hazards, situations and scenarios that the systems need to deal with is critical. In the car industry, it is attempted to define each subpart of the driving process, for example parking. The problem is then limited to only some parts of the operation, for which safety issues are identified.

To identify all the different events that can happen, the analyst must consider an appropriate level of abstraction for the problem, in addition to historical data and experience. The autonomous platform cannot be considered in isolation. The response to an unsafe state is dependent upon the location and environmental conditions. A car, for instance, will operate in different regions, that may have very different traffic patterns (e.g., in Norway / Sweden or in India / Pakistan). As a validation approach, in the car industry, autonomous systems are running in the background while a human driver is controlling the car. This helps to identify new scenarios that the autonomy should react to, but it does not ensure all possible scenarios are covered.

A risk model needs to accommodate the environment, the weather, and the mode of operation. A challenge is the identification of all circumstances that the system may meet. More complex systems may imply that more systematic methods are needed, e.g., to assess the interfaces. Some methods that are currently used and may be applied to autonomous systems include:

- Fault and event tree analysis (FTA & ETA)
- System theoretic process analysis (STPA)
- Functional Resonance Analysis Method (FRAM)
- Simulations

Fault tree and event tree analysis are traditional methods that focus on the graphical representation of the risk analysis. The methods represent events and not every accident is event driven. It may be the circumstances deviating from those expected that lead to an accident.

STPA is a rather new qualitative hazard analysis method. It is a systemic and systematic approach that treats safety as a control problem. It is not limited to component failure, as the more typical risk analysis methods, but it attempts

to identify complex interactive scenarios. Software, hardware, humans, organizations, and regulations can be modeled within the same framework. Also, different levels of abstraction may be employed, and it can be used for all system properties (e.g., safety and security). The main disadvantages are the high need for resources, the lack of competency in the industry, lack of prioritization and ranking, and the lack of the right tools for using the process efficiently.

FRAM is designed to qualitatively communicate risk and the complexity of a system. It is less operational, compared to STPA. The disadvantages are the same as for STPA, i.e., resource intensive, lack of competencies, prioritization, and lack of tools for efficient use.

In general, simulation is a very powerful tool, if applied correctly. It may enable analysts to collect a large number of data for different situations and scenarios at a low cost. Simulation may also include the human, operating in the loop, for a holistic assessment approach. Simulation can be an efficient solution to demonstrate efficiency and transparency of the autonomous system capability.

In the future, simulations should be combined with real (on site) testing to prove to society that the system is safe and capable. It is possible to acquire environmental and operational data offline. Moreover, it is also possible to simulate and test the autonomous systems' responses to very rare events.

Modelling of human, hardware, software and interfaces

Existing risk assessment methods tend to focus mainly on the hardware and human elements of the system. Humans will still be an essential part of autonomous systems in the near future. Depending on the Level of Autonomy, humans will need to remotely control the system or supervise it and step in when problems occur. Models are required to evaluate the contributions from the human in failing to achieve a successful recovery from a problem. It is critical to establish what information is required at handover, how the information is provided, and the time frame for handover.

The probability of software failure is required as input to risk quantification. Software failure is different in nature from physical components. Software will fail when circumstances that have not been predicted by the designers occur or when mistakes have been made by the programmer. Occurrences of these failures are not stochastic but deterministic in nature.

An additional vulnerability in software is that resulting from an intended attack, as stated previously. Hackers exploit unknown vulnerabilities in the system and for critical infrastructure such as transport systems, these may be state sponsored. Since the current and future frequency of these attacks is not

related to their historical occurrence, it is not possible to evaluate this requirement for a risk study. Decisions on the risk posed through such cyber-attacks cannot therefore be evaluated with a risk framework and alternative approaches are needed.

Assessment of self-learning systems

Self-learning autonomous systems may develop their own "personality". They may learn and adapt to specific people and environmental conditions, events and actions. If the analysis of an autonomous system SRS (particularly for the software elements) will rely in testing, updates which change capabilities will need to be formally assessed.

An example is autonomous car driving systems, which exhibit very high complexity. They need to account for all road junction types, regional driving cultures and individual driver characteristics. To physically test a vehicle for all potential options encountered for global operation is not viable. In these circumstances, testing and validation are only possible using simulators which can replicate the full range of options encountered (including those rarely encountered). Simulators can conduct the testing considerably faster than rear time road testing.

Immaturity of risk assessment and validation methodologies in this area pose a potential safety risk. New methods and tools need to be developed. The implementation of autonomous systems should only proceed at the pace of the assessment methods.

Resilience

Resilience engineering approaches were considered to offer an alternative philosophy to risk assessment by which autonomous systems can be assessed and worthy of more detailed consideration. A resilient system is one which can anticipate, absorb, adapt to or rapidly recovered from a disruptive event. This focus on the ability of the system to recover from an unwanted event gives a means by which software malfunction may be evaluated without the need to predict the occurrence frequency.

When a system fault is observed, a response needs to be fast and the initial incident management may have to be performed without knowledge of its cause. Once the cause is determined a transition from incident management to full system rectification can be implemented. Such an approach enables the system to be safely operated in all circumstances, not only those with a low risk prediction.

The potential benefits of such a resilience approach needs further investigation. Measures of system performance (MoP), which should be predicted throughout any incident, would need to be established. It is expected that these will vary depending on the occurrence of a safety problem or reliability problem. Methodologies to predict how this MoP varied through the phases of threat occurrence, system performance degradation, incident management and full system recovery would be needed and the exact definition of resilience which was predicted from these factors established. Different MoPs would be required for different autonomous system applications.

Real time operational decision support

Conventional risk analysis techniques such as fault tree analysis are usually used off-line in order to certify that a particular system delivers acceptable safety performance. An autonomous system will need to establish when it is no longer operating safely and requires a mission abort strategy to be activated. Determining unacceptable performance can be rule based or it can exploit the system failure analysis approaches in real time to predict when the safety performance is no longer acceptable. Events which represent deteriorated or failed hardware (established through fault diagnostics), changes in environmental or operational conditions can be input as updated event probabilities to the system failure models. The analysis of models formulated as a fault tree can be rapidly performed using Binary Decision Diagrams.

Such approaches have been explored to establish unsafe conditions for pilot-less aircraft, UAVs (Unmanned aerial vehicles), the timeframes in which decisions need to be made would certainly make these approaches applicable in the maritime and marine applications. Since the operating environment of aircraft is less complex than cars, the response time of such predictions may currently limit the potential for automobile application.

Data requirements

Several types of data are needed for the SRS assessment of autonomous systems, including

- Sensor data and understanding of their usage
- Service, repair, warranty and maintenance data
- Experimental data and test data
- Condition data
- Surrogate data gained through simulations
- Data on the frequency and nature of cyber-attacks on the system

These data may be used in real time, in virtual and dynamic models, to manage failures and plan and predict maintenance. However, the data needed is frequently not available and, when available, may be of low. Standards for data collection are needed across companies and sectors.

Some data may be transferred between industries. For example, information related to human factors and human error may be transferred between highly automated systems. It is important that historic data, or data from manned systems, is assessed for their applicability for autonomous systems. Similarly, environmental data needs to be assessed for case relevance.

Data needs to be analyzed together with the associated uncertainty, to ascertain if data is complete or if there gaps in the observations, due to an insufficient monitoring frequency.

Conclusions

From the group discussions the following conclusions were drawn regarding the safety, reliability and security of autonomous systems.

- The software elements of autonomous systems challenge the applicability of current risk assessment approaches. This is due to software malfunction being very different from hardware or human failure and not stochastic in nature. Since their historical occurrence does not indicate future expectations it is not possible to formulate their expected likelihood or frequency. The same problem exists in predicting the frequency of malicious, intentional attacks on the software. Self–learning features of the software also add difficulty in the validation of acceptable performance.
- Many of the currently available methods can still play a part in supporting the safety, reliability and security of autonomous systems.
- New modelling techniques, which holistically capture the strong connectivity and interdependencies between software, hardware and human operators are required.
- Simulations provide a practical approach to assist in the detailed understanding autonomous systems with respect to SRS, to collect data, and to validate systems SRS behavior.
- The concept of resilience engineering is an alternative approach to risk assessment and offers a focus on absorbing and recovering from failure events which can be applied without knowing the frequency of the failure.
- Data requirements are application specific with standards required to ensure quality.

- Quantification methods for SRS can play a bigger part in the future. In addition to the certification process for an autonomous system they can be incorporated for real-time decision support during a mission to identify when the system performance drops below the acceptable threshold.

Group Participants

John Andrews
University of Nottingham, United Kingdom

Jonas Borg
Volvo Penta, Sweden

Kenneth Titlestad
Sopra Steria, Norway

Markus Heimdal
Rolls Royce Marine, Norway

Matthew Minxiang Hu
Haylion Technologies, China

Nikolaos P. Ventikos
National Technical University of Athens, Greece

Odd Ivar Haugen
DNV GL, Norway

Osiris Valdez Banda
Aalto University, Finland

Siv Randi Hjørungnes
Rolls Royce Marine, Norway

Stein Haugen
Department of Marine Technology, NTNU, Norway

Thomas Johansen
Department of Marine Technology, NTNU, Norway

Torgeir Moan
Department of Marine Technology, NTNU, Norway

Yan-Fu Li
Tsinghua University, China

System Verification, Processes and Testing

Report of the Discussions of Breakout Session

Authors: Tristan Perez (Session chair), Andrey Morozov, Børge Rokseth, Jon Arne Glomsrud, Matthew Luckcuck, Thor Myklebust, Tobias Torben, Xue Yang

A challenge related to autonomous systems concern their verification process and testing. This discussion is not detached from regulatory, societal, and ethical requirements. Indeed, being able to verify issues of governance and ethics is of high importance; yet, a key concern is which governance body and whose ethics are being adopted. The verification process should not be entirely removed from these concerns, and ensuring that the right properties are being verified will require interaction with domain experts in those areas. The regulatory, societal, and ethical requirements should be included at the beginning of the design process and should be fed through to the verification phase. However, the verification process may identify ethical concerns (especially if they have not been identified during the requirements and design process) and engineering practice should ensure that these concerns are included into the system's design.

A main concern is how to obtain the right requirements against which to verify the system. While the validation of requirements is a concern with the verification of any system, it may be a particular challenge with autonomous systems. Firstly, this is because of the complexity of autonomous systems; secondly, this is because of a lack of consensus on regulation and ethical guidelines for autonomous systems.

An additional concern is the identification of the best verification processes to use for autonomous systems. Given their complexity, their embodiment in the real world, and their potential for adaptation or learning, continuous and integrated processes are recommended. Briefly, the adoption of a more DevOps-like approach and designing online (continuous or periodic) re-verification systems.

Other elements of the discussion on the topic includes methods for communicating the results of verification efforts and the inclusion of formal methods. Communicating verification efforts to both regulators and the public is important to ensure autonomous systems can be certified, by a regulator, and trusted, by the public. The application of formal methods to the development of autonomous systems can provide automatic verification and unambiguous specification of the system's intended behavior. However, how the autonomy is

implemented can have an impact on how challenging the application of formal verification can be.

The following sections deepen those discussions.

Verification Processes for Autonomous Systems

In general, the verification process for autonomous systems should contain firstly an initial verification and testing process and secondly, an on-going process to deal with changes in the system or its operating environment. This can be achieved by adapting classical models such as the V-model into DevOps-like models.

Considering behaviors are the key pathway towards frameworks for certification of autonomous systems. Behavior can be evaluated in terms of safety, performance and ethics. The process of initial verification and testing consists of first identifying desired behaviors for safety, performance, security and ethics. Then metrics and verification criteria must be established before the actual verification and testing activities take place. After being built, verified, and accepted a system may change due to, for example, software updates or any potential learning ability of the system. In addition, the system environment may change. For example, an autonomous car may be taken to a new area where other cars and pedestrians behave differently. The on-going verification process is intended to deal with such changes. In order to achieve this, changes must be detected and analyzed to determine the effect in terms of verification needs. The verification and testing process can be discussed in terms of three steps:

1. Defining desired behavior for the autonomous system;
2. Identifying and conducting tests and verification to satisfy verification criteria;
3. Monitoring systems and conducting change analysis during operation to detect any new needs for verification due to system or environmental changes.

Step 1

When defining the desired behavior for the autonomous system, it is necessary to determine a level of granularity at which the desired situational behaviors are defined. This raises questions, such as, to which levels systems should be decomposed and how systems should be decomposed. In general, desired behavior should be specified at the system level and then refined as much as necessary into components or sub-functions in order to determine sub-system or sub-function behavior that ensures the desired system level behavior.

Methods for specifying desired behavior may need to be specified case by case. It is reasonable to assume, however, that elements of hazard analysis (i.e. identifying what can go wrong and how the system should handle these situations) as well as formalizing design requirements and requirements from standards into behavioral models will be highly relevant approaches. An important part of the documentation of this step will be to record the assumptions made regarding the system and its operating environment.

Step 2

The next step of the verification process for autonomous systems, is to identify and conduct tests and verification to satisfy verification criteria. The goal of this step is to observe the system behavior under tests and other verification activities, and to evaluate the observed behavior to determine our confidence that the system will behave according to the desired behavior. Increasing this confidence corresponds to reducing uncertainty. There are two general types of uncertainty in this context. First, there is uncertainty about whether an observed behavior should be classified as desired or undesired behavior, and secondly, there is uncertainty when a certain behavior is observed in one scenario related to the extent to which this can be considered representative for similar scenarios. Verification then is about collecting evidence to reduce these uncertainties. To achieve this, verification needs both to be broad in terms of capturing as many types of scenarios as possible while it also is necessary to test each type of scenario extensively to ensure that results are representative of all similar scenarios. A formal verification, model-in-the-loop, process-in-the-loop and hardware-in-the-loop methods may be central methods for collecting evidence to reduce uncertainty and increase confidence.

Verification of autonomous systems may be more resource demanding than verification of traditional systems because there will be more focus on system behavior and there can be a huge number of possible behaviors. It will be more critical for autonomous systems than for human operated systems to foresee abnormal scenarios because the autonomous systems may be less robust and innovative with respect to handling the unforeseen. Therefore, any possible scenarios must be foreseen and considered in the verification process. This may cause state explosions and the necessity for rare event simulations.

Step 3

The third step of verification is to monitor operations and detect emerging verification needs during the operational phase of the system. One important aspect of this is to define the operational environment for which the system has been verified, as well as the system that has been verified. The assumptions being made regarding the system and its operational environment must hold true for

each conclusion reached during verification to be a valid verification. Once these assumptions are known they can be monitored during the operational phase of the system and if one no longer holds true, further verification is necessary. Change analysis is proposed to achieve this.

Communicating Verification Results

The group discussed the challenge of communicating the results of verification to stakeholders –regulators and the public.

Some sectors in which autonomous systems are being explored require that a system is certified. Regulators need to be able to understand how an autonomous system is verified (and be confident in the verification results) in order to certify a system for use. Given the complexity of autonomous systems and their potential to change (either through learning, self-reconfiguration, or simply by changing their operational environment) efforts must be made to ensure that verification approaches for autonomous systems are amenable to the regulator(s) of the sector in which they are to be deployed.

Communicating the concept and results of verification to public is key to gaining public trust of autonomous systems. Society seems to have lower tolerance for accidents and unexpected behavior from autonomous systems, so efforts to ensure public trust should help with the adoption of autonomous systems in meaningful use cases within society. Results from verification must be interpreted and presented in a way that helps decision-making, such as, whether it is safe to deploy a system into society. Other key challenges here are how to communicate the level of confidence and uncertainty in the system's ability to continue to operate according to desired behavior, and how to communicate what the desired in a digestible way.

Formal Methods

Formal methods are mathematically defined techniques to the specification, design, and verification of computer systems and software. They enable the expression of requirements and description of systems with precision and no ambiguity. Often the tool support for checking that a system exhibits the required properties is automatic and exhaustive. The formal specification and verification of autonomous robotic systems is an ongoing topic of research for the formal methods community.

The successful application of formal methods to autonomous systems can largely depend on how the autonomy is implemented. Neural networks, for example, are challenging for formal methods to deal with because it is often not

understood how they produce their output. Formal methods work best with more symbolic approaches to autonomy.

Arguments were put forth arose that including formal methods into the specification, design, and verification of autonomous systems is very important because of their increasingly safety-critical nature. Formal Methods can be introduced at several stages during the development process. For specification, they can help to clarify the requirements (and even check that the requirements themselves have not introduced unintended errors). During design, they can be used to check that the designs meet the requirements. During verification, various automatic tools exist to exhaustively check that the description of the system preserves the required (safety, legal, ethical, etc.) properties. This automation will help with the DevOps-like process of ongoing verification described above.

An obvious final challenge is that of ensuring that the final system represents the formal descriptions of the system, and so preserves the required properties. This is a challenge faced by any software development process. Some formal methods can verify program code (for example the Agent Java Pathfinder, a program model checker for agent-based autonomous systems) and there are other methods from which program code can be automatically generated. Even without these types of method, using formal methods during the requirements and design phases can help to reduce errors introduced at these early stages of the development process.

Conclusion

Six main challenges and four distinct opportunities related to verification and testing of autonomous systems can be pointed out. The following challenges were identified:

1. The V-model may no longer be adequate and is necessary to either replace it or adapt it into a DevOps-like model.
2. Autonomous systems may sometimes need assistance from operators, and in certain scenarios, control needs to be handed over from the autonomous system to the operator. Verification of the control handover may be a particular challenge
3. In traditional systems, the behavior of the system is to a greater extent governed by human operators than what will be the case for autonomous systems. Operators are often trained and certified, and together with their general human experience. This is accepted as sufficient. Once the system behavior starts being governed by software, rather than human operators, how does this process

translate to training and certification of human operators and the consequent level of trust?

4. Learning algorithms may be central to autonomous systems control. A specific verification challenge is how can trust in a system be established that may continue to adapt itself after deployment. Thinking of the verification process as ongoing through the life cycle of a system will be a central issue with respect to this challenge.
5. It will also be a challenge to formulate and parametrize desired behaviors. It may be close to impossible to cover all operational profiles. While systems operated by humans have a certain robustness because they can adapt to situations, and as such can handle unforeseen scenarios, autonomous systems are not robust in this sense. This means that any scenario must be foreseen, and a system response must have been planned for the system to be able to handle this situation.
6. In order to cope with a huge number of scenarios, automated and customizable methods and tools for verification and testing must be developed.

While there are challenges related to verification of software rather than human operators, who are governing the behavior of systems, there are also opportunities related to this. In addition to the six challenges, four main verification and testing opportunities for autonomous systems are identified:

1. When the human operator is replaced by software, this enables replacing periodic inspections with continuous performance monitoring which can be used to revoke operating license in the event of inadequate performance.
2. The behavior of software can be considered more deterministic compared to human operators. In general, it is believed that it is possible to predict the behavior of software with higher precision than that of human operators. While it is not possible to inspect an operator's brain to determine how the operator will respond to different inputs, it is possible to inspect the software code to determine this.
3. Once the human operator is out of the loop, it is possible to predict and verify behavior online through online model-based verification where variations of the current operational scenario can be simulated into the future to verify safe system response.

4. With human operators in the loop, automated accelerated testing of the complete system is not possible. With the human out of the loop, testing can be conducted in simulators faster than real-time.

Group Participants

Andrey Morozov
Technical University, Dresden, Germany

Børge Rokseth
Department of Marine Technology, NTNU, Norway

Jon Arne Glomsrud
DNV GL, Norway

Matthew Luckcuck
University of Liverpool, United Kingdom

Thor Myklebust
SINTEF Digital, Norway

Tobias Torben
Department of Marine Technology, NTNU, Norway

Tristan Perez
Boeing Research and Technology, Australia

Xue Yang
Department of Marine Technology, NTNU, Norway

Intelligence and Decision Support

Report of the Discussions of Breakout Session

Authors: Christoph A. Thieme, Marilia A. Ramos

Autonomous systems and Artificial Intelligence

Autonomous systems are cyber physical systems. They may include Artificial Intelligence (AI), e.g., in decision making or other autonomous processes. However, although an overlap exists, an autonomous system is not a subset of AI, nor vice versa. Reasoning is an important aspect of autonomous systems and AI. Reasoning refers to the ability to explain actions and decisions. For humans, many actions are learned intuitively and do not result from reasoning. Still, the reasons for these actions can be described in retrospect, even though it was based on intuition.

AI needs to be interpreted in a simplified manner than what is currently expected in the public opinion from AI. AI is a collection of mathematical methods, helpful for solving tasks associated with intelligence. AI methods try to find regularities in sets of data. An operational perspective may help to better clarify the concept: AI is mainly comprised of some form of learning, some degree of reasoning, interaction with an environment. It should have an ability to explain how or why the AI made its internal decisions.

The definition that "AI is always the thing that humans can do and machines cannot do" is not suitable and in itself unachievable. However, AI has advantages over humans, for example, analyzing quickly large sets of data. Therefore, the human standard might not be ideal. AI may be used to learn the operational parameters for an autonomous system, identify weights for risk factors, or detect abnormalities.

The focus of AI should lie on the autonomous system, meaning that AI methods comprise tools that may help to realize autonomous systems. Autonomous systems are more than AI, since they comprise hardware and other software. By excluding the physical systems from an autonomous system and reducing it to AI, the extended Turing test is not achievable, i.e., one is unable to detect different behavior of AI and humans.

Adaptive autonomy is an often-used term in the context of autonomy and AI. However, adaptive autonomy is an ambiguous term. It may refer to a system that uses a learning (AI) system, to a system that changes the degree of autonomy during an operation, to software updates that adapt the system when needed, or

to the behavior that occurs adapting to a situation. Commonly, self-adaptive systems change their behavior based on experience.

Risk in control and decision making

For intelligent autonomous systems it is required for risk to be considered during the early design phases. Decisions of an autonomous systems need to be based on an implementation of risk considerations that are defined clearly mathematically and operationally. Risk is often defined in probabilistic terms, in a pseudo mathematical equation: risk equals probability or frequency times consequences. Successful implementations of (quantitative) risk assessments (QRA) in applied projects on autonomous systems should be developed to highlight the advantages of QRA.

AI may be seen as a factor contributing to risk. However, it is generally the complexity of the system from which risks emerge. Hence, it is important to understand the system, not only the AI. Systems might become so complex that nobody can see the full picture. Therefore, it may be impossible to understand the associated risks and failures that may occur. Methods and approaches are needed to manage and assess complex systems.

There is no essential difference between decision-making and optimization based on parameters. Autonomous systems and AI algorithms make continues choices between a spectrum of operational parameters. Making only discrete choices is not a property of an autonomous system, besides decision making; parameters are optimized to achieve the most efficient execution under the given circumstances. This resembles the behavior of, e.g., human drivers that follow a set of rules and optimizes constantly the vehicle speed and heading, and their own behavior to avoid accidents and penalties.

Therefore, risk is a cost and a constraint for operation of an autonomous system. Different types of risks emerge for autonomous systems that are not possible to be covered by only one measure to measure the risk level of operation. Identification of relevant risk measures and risk factors is one of the main tasks during development. These need to be implemented in the control system and the decision module of the system.

During the development of an autonomous system, a baseline performance needs to be defined, as reference for acceptable performance and risk. The baseline performance should not be lower than the performance by a human operator. One challenge that arises when approaching performance requirements is that the evaluation of the human performance is difficult. The performance acceptance criteria may be vague. Perceived risk versus real risk is

also relevant with respect to this evaluation. For example, car crashes occur frequently, while an autonomous vehicle accident is paid much more attention and is perceived as more severe by the public.

Risk reduction across the system architecture

Several definitions, views and hierarchies exist for control systems. The system architecture depends on system purpose, size and complexity. A general guideline is that low levels of control are reactive, while higher levels of control are more proactive. Higher levels may include a proactive planning layer and a supervisory layer for fault detection. Using the term "executive layer" may not be useful, since everything is executed.

In each of these three suggested layers, different risk measures may be used. This depends on the layer purpose and the anticipated level of autonomy. Several risk measures may apply, e.g., probability of failure and probability of collision, mission failure, system failure. Using one measure across all levels is not sufficient. Current systems do not include explicit risk models in their control structure.

One challenge for the supervisory layer is the identification of the fault source when a fault is detected. Subsystem integration between components and systems may be inadequate and detected faults may be propagated from the real source. A clear structure and hierarchy are needed to filter faults and identify their sources.

The risk that emerges during an operation may be reduced prior to operation and during operation (post-deployment). However, risk reduction should be mainly achieved during the design and pre-deployment phases. Risk reduction may be achieved in all architectural levels.

In the post-deployment phases, the risk level and the system condition need to be monitored. The autonomous system should detect critical and pre-critical situations. It should use pre-critical situations to avoid critical situations in all equivalent systems. For identification of such situations, the autonomous system may use statistics or other machine learning (ML) approaches.

Development of safe autonomous systems

Industry practice shows that risk assessments and modeling are necessary processes in the development of autonomous systems. Using risk-informed decisions enables better design decisions. Through integration of the risk information, it is possible to identify opportunities for monitoring and prognosis of failure development. This in consequence will reduce the need for unnecessary maintenance. ML algorithms may be one tool to monitor the system. There are

"best practices" in the software industry for testing, validation and verification. However, there is not a general recipe for future developments. Simple, "if-then-else" structures, can be proven to work correctly and reliable. For ML and AI learning techniques, these methods are not available yet.

Risk models need to reflect the assumptions made in the system design. During design, these assumptions need to be identified and it must be consequently assessed how the level of risk may be affected by these assumptions. So-called legacy systems, systems that build on former generations, build on certain inherited assumptions. However, it is often undocumented why the assumptions were made. In a few cases, it is assessed if the assumptions are still reasonable.

Currently, airline pilots report anomalies based on previous experience and training. A system should be required to self-report data that can be used for further development and improvement. Near misses are a significant learning source for autonomous systems and AI. They provide more insights than just the accidents themselves.

Self-adaptive systems must be designed to detect if the adaptive behavior is performing worse than the previous learned behavior. Mechanisms need to be in place to return to proven and safe behavior in such cases. The most safety-critical parts of the system should not build on adaptive methods. A predictable and verifiable behavior is required. It is necessary to define what changes need to be verified from the outside and which can be done based on learning.

A hierarchal structure regarding the regulation of autonomous systems is required, analogous to the regulatory framework for current human operators. Since autonomous systems will become a reality relatively soon, the regulations need to be put in place. However, there needs to be room for future improvement and adaptation. The autonomous systems will not appear abruptly, and systems will change incremental. Certification for drones, for example, has requirements in place, to be commercially viable already. For consumer drones such rules are, for example, the inability to fly into no-fly zones, etc. Newly emerging companies, working on autonomous systems may be less conservative than the established companies. Hence, regulations are needed to create a common baseline.

Risk awareness in autonomous systems

Decision making

Improved intelligence and online decision-making capabilities are needed in autonomous systems. Existing control theoretic approaches are not explicitly connected to risk assessment and modelling. Some control strategies use

methods that deal with constraints and unwanted states. This leads to robust control but tends to be conservative. A clear risk definition is needed for that purpose. Risk consideration should also include events that are not known. Simple control strategies and models cannot include such considerations. There exist few control strategies for handling extreme cases with low probabilities.

There is another gap between control theory and control practice today. Switching between discrete states is used to adapt to certain situations. There is a lack of usage of established control strategies in practice. Proactive approaches are required: Actions in case something might happen, being ready for "black swans", i.e., rare but extreme incidents. In contrast, a reactive approach would imply to act when something is happening. In any case, there is a difference between safe behavior and safe state. In certain situations, it may not be practical or safe to go into a safe state, e.g., shut down the system, or stop it. A safe alternative needs to be designed and chosen.

Model predictive control (MPC) is one control strategy that is suitable for autonomous systems. However, using only one risk measure in such a control strategy is not suitable. A vector of several risks is needed to optimally use the method; these may be probability of collision, time to collision, etc. Risk may be then a cost and a constraint in the MPC algorithm. Using risk just as an optimization criterion for minimization would lead to the system never starting, since then the risk is lowest. In addition, using the risk as a constraint enables the user to demonstrate that the system will not accept a higher risk than prescribed by a legislator. In the MPC method, this may have the disadvantage, that the system will always choose a solution close to the accepted risk limit.

Online risk models may assist in decision-making. Online risk models are models that have been developed before the mission is executed and that use data measured online to constantly update the current risk level. Necessary data measurements can be identified in risk analysis. It may be possible to sample the measures directly or it may be necessary to use risk indicators. ML may be used to tune the different risk factors and other objectives to give the behavior we want. Game theory may be useful, too. It must be taken into consideration how other entities involved might act.

An intelligent system must not only follow the rules and trust that other traffic participants do the same. An autonomous system must be able to detect or predict intentions. A good example is the maritime sector, where COLREG rules exist. However, human navigators may violate these. Initially in the aviation traffic collision avoidance system (TCAS), for example, only positions were communicated. This was not an intelligent system, since it only detected other planes, but did not coordinate maneuvers with each other. After serious

accidents, the rules of behavior had to be adapted and the system is now more prescriptive and solving traffic situations automatically.

Health monitoring

The performance of an autonomous system also affects the decision possibilities. Hence, it is necessary for an autonomous system to be aware of its health status. Parameters that define the health status are the conditions of sensors, actuators and the control system performance. In addition, mission external parameters, such as information on maintenance and available spare parts in the operation basis may affect the decision possibilities for an autonomous system.

Two different time horizons need to be taken into account with respect to health monitoring: the long time perspective gives information on degradation of components the overall system's condition that may be used for service planning, e.g., changing parts, and general maintenance. The short time perspective provides information on the system's performance degradation, its effect on the mission outcome, and the ability to handle possibly critical situations.

For a system to detect that its performance is degrading, it needs to be designed knowing the baseline performance. Risk assessments are essentially identifying what types of situations the system cannot deal with. Therefore, risk assessments aid to identify performance requirements to the design of the systems and the operational design limitations. The system is then limited to function properly in situations the designer managed to envision. Hence, the system also needs to be able to detect that it is operating outside the operational design envelope.

Input for this type of behavior needs to be supplied in manuals for sensing equipment. Similar to the commonly found curves "efficiency vs. environmental parameters", the measurement uncertainty could be described by the behavior over a combination of environmental parameters. However, a device may not be tested in all operational conditions. Then the reliability data needs to be produced by the user, e. g., NASA is producing reliability information for most of the components themselves, such as charge and discharge curves of batteries under extreme temperature conditions.

Sensors may be subject not only to physical degradation, but also to snow, fog, dust, or alike. This may inhibit the performance. In addition to monitoring the physical condition of the system, information needs to be combined and the reasons for degraded performance need to be detected. AI methods may assist to monitor the system health and detect the causes to a degraded component.

Sensor requirements

Sensors need to be reliable during an operation. The environment influences the uncertainty of measurements. Components degrade, which increases the uncertainty. An autonomous system needs to handle these facts by sensor redundancy, better sensors, etc. However, redundancy adds to costs, increases power consumption, and adds weight. One possible solution could be using the payload sensors as redundant sensors (e.g., using visual flow to validate accelerometer measurements). The ability to handle uncertain situations is also needed when facing sensor degradation.

An autonomous system is not only about sensors, but it must be able to comprehend the meaning of the sensor measurements. Many systems can detect if the weather and climate conditions exceed the design limitation.

Data requirements for safety and reliability analysis and safety monitoring

Data and information from the sensors should be reliable and available when needed. In addition to pure measurements, sensors should give information on the sensors' uncertainty of the measurement. In this way, the uncertainty and the effect on the system may be assessed mathematically.

It is known that navigation systems are prone to both noise and design flaws, which may be undetected until their effect is experienced. However, when the system is deployed, it may be too late to correct the error. Hence, an appropriate design process needs to be chosen, to ensure that necessary data is collected in the appropriate frequency and quality.

One approach may be to use information trees, which are similar to fault trees. The challenge with such a tree structure is the interpretation of the Boolean logic. The trees can be used to identify:

- What is the information that needs to be gathered (the top event)?
- What needs to be measured based on what the needed information?
- What types of data and sensors are needed to meet the knowledge condition in the top event?
- Which data types are dependent or independent?
- What are the success metrics?
- Where are the best places to collect information?

Internal and external data uncertainty

Three types of uncertainty can be differentiated:

1. Measurement or data uncertainty
2. Model uncertainty
3. Interaction uncertainties

Measurement uncertainties are well defined and inherent to the measurement system and method. Methods for describing this uncertainty are well established, e.g., Gaussian distributions. The uncertainty can be expressed numerically. Its effect can be propagated through the system and the effect can be assessed. Sensors should give information on the certainty of the measurements.

Model uncertainty reflects the completeness and correctness underlying the models that are used in a system. Statistical distributions may not be able to capture this type of uncertainty and some parameters that are used in a model may be highly uncertain, e.g., turbulence is difficult to capture numerically. Assumptions need to be made that are imperfect. Model uncertainty may be introduced to keep the system simple. Adding many parameters, whose effect is uncertain, will not improve the model. Hence, parameters may be neglected, if the effect is highly uncertain or negligible, in order to make the system more efficient.

The third type of uncertainty is the uncertainty with respect to the interaction with other parties, humans or manually operated/autonomous systems. The behavior of others is difficult to predict. An autonomous system may be “perfect” in itself. However, other traffic participants may cause an accident.

The system is a conservative system if the estimated uncertainty exceeds the real uncertainty. Unsafe behavior is to be expected if the estimated uncertainty is lower than the real uncertainty in a given situation. Risk analysis is required to estimate the uncertainties with respect to the control environment. The analysis needs to include the operators or supervisors and the autonomous system themselves. The state of the operator needs to be reflected in the control system.

With respect to the third type of uncertainty, one challenge is to robustly detect and identify obstacles and other participants. For AI methods, such ML and deep learning, it is difficult to predict their output, due to their prediction accuracy and often in tracible behavior. Both, identification and prediction are time dependent. Small variations in timing may affect the predictions. A test approach may require a very low uncertainty level, which will correspondingly take a lot of time. Without a verifiable equation, it is difficult to quantify this

uncertainty. There are methods for handling noise, i.e., environmental disturbances, but the theory is lacking when it comes to working with probability density functions.

An approach to verification should be to test first the algorithms, e.g., through simulations. These tests, then need to be validated in the real operational environment. There is need for clear guidelines and checklists for building an autonomous system.

If an operator or supervisor is involved in an autonomous operation, it may be necessary to monitor the operator and assess the uncertainty with respect to the operator's ability to cope with a situation. Information could be extracted from the performance during the current task and projected on the execution of the next task. The visualization of uncertainty to the supervisor/ operator remains one challenge.

A core demand for AI-based systems is that they need to be able to detect if they are outside of their operating range. The system needs to detect if the uncertainty in a given situation is too high. This includes the detection of anomalies that were not included in the training data sets and the appropriate reaction to these. This can be compared to a human driver who will adapt to a new situation and identify untrained situations.

Autonomous systems´ interaction with the operator

Autonomy in many cases shall reduce the number of operators needed to run a system safely. There are benefits to be realized with higher levels of autonomy. A system does not need to be fully autonomous to be cost effective. There are different areas or operational time intervals, where it is better to be more autonomous, e.g., for the maritime industry at open sea. In any case, there will be an interaction with humans even for fully or highly autonomous systems.

In many planned autonomous systems, the operator takes a supervisory role. The operator is used as backup to cope with situations that the system may not be able to handle. When such a situation is detected, the level of autonomy can be changed. It is critical that the communication between autonomous system and pilot is adequate. Information needs to be presented clearly and comprehensively.

The operator needs to know the state of the vehicle when receiving control. There should be a smooth transition between autonomous piloting and human piloting. It is important to identify the necessary information for the operator to carry out the necessary actions. The system needs to be designed accordingly. Recent accidents in the aviation industry show that pilots need to be

trained sufficiently in order to not fight against the autonomous systems. It needs to be defined what is part of the "autonomy" and what is the human's role.

The state of the human operator must be taken into consideration when attempting to give control. The workload for the operator may increase and it may be safer to continue autonomously, as there might not be enough time for the human to react or if the human may be unable to react. For an unmanned aircraft, it is better to use the autonomous system when taking off and landing, because of the increased stress levels and reduced perception capabilities of the operators during these tasks.

One concern is that the human operator suffers from skill degradation over time without continuous training. The operator may also suffer from a low workload and decrease of situational awareness. Similarly, one aspect that needs to be assessed in design is the confusion by sudden error messages to the operator, so called automation etiquette. The design of warning and handover messages needs to be clear. The autonomous system cannot just be stopped in the middle of the operation, e.g., this may create hazards for other participants. It is critical that the autonomous system relies on the operator when operating outside the design envelope rather than in a predefined set of situations that may be actually manageable by the autonomous system.

Industries that are currently attempting to automate their systems and products, such as, the automobile and maritime, must learn from aviation. Especially, skill degradation is widely researched in this field. Just assuming that the human is a good backup when the autonomous system reaches its operational limitations, is not viable.

Autonomous systems interaction with each other/ other systems

To improve cooperation and the predictability of the behavior of autonomous systems communication of planned actions is needed. Consequently, communication standards are necessary to be developed. In the future, an autonomous system might communicate with an infrastructure to get high-resolution maps or similar information about the area or attain feed forward information from a non-autonomous agent. It may enable people and other systems to better understand the current state than just by looking at the current behavior. Communication may also reduce time-delays, which is especially relevant for slow responding systems, such as ships.

Non-autonomous systems may benefit from using the information on the future actions and intentions of an autonomous system. For example, in the

maritime sector, information on pilots´ actions (rotating the steering wheel) could be fed forwarded to the autonomous system and communicated to other systems nearby, such that these do not need to detect that the ship is turning.

Group Participants

Adrian Arjonilla
UAS Consulting, USA

Arne Ulrik Bindingsbø
Equinor, Norway

Edmund Brekke
Department of Engineering
Cybernetics, NTNU, Norway

Ole Jakob Mengshoel
Department of Computer Science,
NTNU, Norway

Rudolf Mester
Norwegian AI Lab, NTNU, Norway

Sebastien Gros
Department of Engineering
Cybernetics, NTNU, Norway

Simon Blindheim
Department of Engineering
Cybernetics, NTNU, Norway

Sverre Rothmund
Department of Engineering
Cybernetics, NTNU, Norway

Sverre Torben
Rolls Royce Marine, Norway

Tarannom Parhizkar
Department of Marine Technology,
NTNU, Norway

Tom Mace
UAS Consulting, USA

Tor Arne Johansen
Department of Engineering
Cybernetics, NTNU, Norway

THIS PAGE INTENTIONALLY LEFT BLANK

Papers

Human-System Interaction in Autonomy Method – a Structured Approach to Risk Monitoring

Marilia A. Ramos(*); Christoph Thieme

Department of Marine Technology, Norwegian University of Science and Technology - NTNU, Norway
(*) marilia.a.ramos@ntnu.no

Introduction

Research and development projects on autonomous systems have faced increasing interest, and some are currently in a testing phase. Autonomous systems' operation may be safer than traditional manned systems, since human error may be a contributing factor to many accidents. Nevertheless, a fully autonomous systems with no supervision and/or interference from humans are not expected soon. The operation will thus rely on a human-autonomous system (H-AS) collaboration. This interaction may not be constantly the same and the role and tasks of the operator may change. Then the autonomous system is designed with a dynamic Level of Autonomy (LoA), i.e., the LoA may change during operation depending on certain conditions.

As humans will still be involved in the operation at some level, human error may still occur [1–3]. In addition to human error, autonomous systems create new challenges, such as increased cyber security threats, detection of unforeseen conditions and actions from other people or the possibility of losing communication with other partners. Hence, risk assessments of operation are important [4]. They face two main challenges: i) the strong reliance on H-AS collaboration during the operation, and ii) the possibility of a dynamic LoA.

Few publications address topics related to hazards and risks associated with autonomous systems' operation. A recent review [4] of risk models aiming conventional and maritime autonomous surface ships (MASS) revealed that current approaches do not sufficiently model the functions carried out by software-based systems and that human operators are often treated superficially. Different operational modes of vessels are only covered to a limited extent. The current literature concerning autonomous systems does not model and analyse the H-AS interaction as potential contributor to the risk of operation, nor does it reflect the dynamic LoA of the operation. The Human-System interaction in

Autonomy (H-SIA) method intends to fill this gap. The method, although being developed foremost for MASS, is generic in nature, reproducible and structured.

This paper summarizes the H-SIA method, its background advantages and current limitations. More detailed information on the method and a case application can be found in the full article [5].

Methodology

The H-SIA method, presented in this Section, is initially composed of two elements: (i) an event sequence diagram (ESD), and (ii) a concurrent task analysis (CoTA). The method was specifically developed for and applied to collision scenarios between an autonomous ship and another vessel or object. Nevertheless, it is expected to have general applicability for autonomous systems.

Figure 1 presents the three main steps in the H-SIA method. Steps 2 and 3 are described in more detail in the following sub-sections. The general approach comprises familiarization (Step 1) to ensure that the analyst can apply the flowchart for the ESD development. The ESD development is the second step, where the ESD is built by answering design related questions of the autonomous system and the LoA of its operation. The developed ESD can be further analyzed with the CoTA (Step 3).

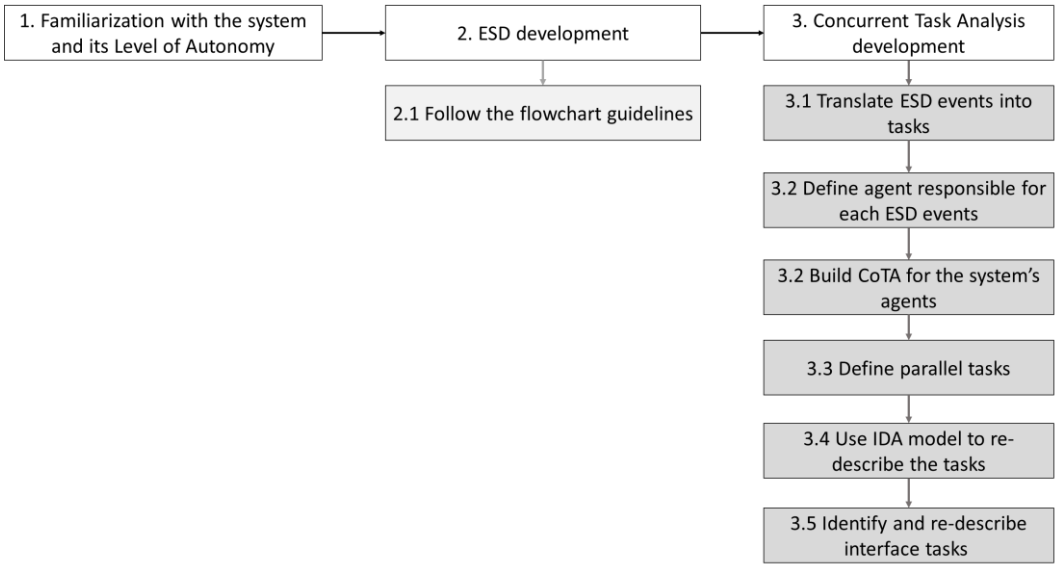


Figure 1: H-SIA method application steps (from *[5]*)

Figure 2 presents a general view of the H-SIA method results. The CoTA is success-oriented; it describes the tasks involved in the success paths of the events of the ESD. The interactions between the interface tasks of the agents are

indicated with circles: a circle with an arrow exiting the event indicates that the task results in an output necessary to the accomplishment of a specific task of the other agent. Similarly, an arrow entering the event indicates a task that receives input from a specific task from the other agent. Interactions are identified by following the rules for task re-description and the CoTA stop rules. The events in the ESD cover either events related to the human operator or the autonomous ship. Some events may be related to both entities.

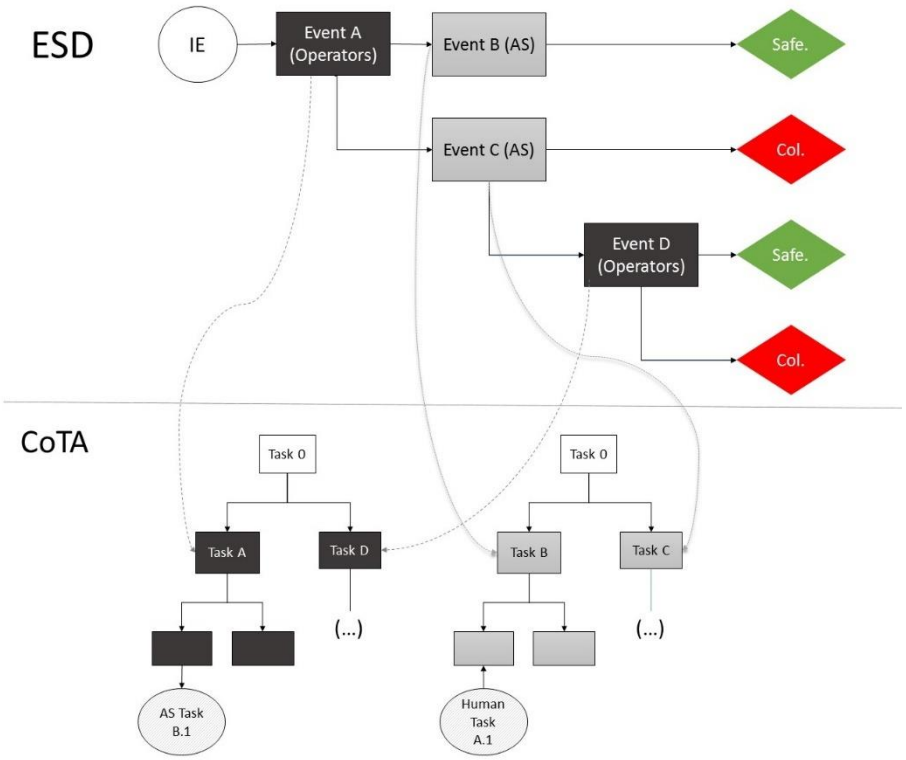


Figure 2: Simplified example of H-SIA method elements. (Adapted from *[5]*) Abbreviation: AS - Autonomous ship

Event sequence diagram and flowchart for development

ESDs are a generalized form of event trees. The ESD framework is flexible in modeling the behavior of key processes and hardware and operator state changes. The timing aspect is considered through the order of events. Thus, it is a more literal representation of a system state than event trees [6]. ESD are used, e.g., in the Phoenix Human Reliability methodology, which makes use of a flowchart approach to build a Crew Response Tree [7–9]. This is encouraging to apply the ESD framework and flowcharts for their development.

H-SIA provides a flowchart for the ESD development. The questions guide the building of the ESD and assist in including only relevant issues in the ESD that appear in the logic order of the questions. The use of the flowchart ensures

traceability and reproducibility of the analysis. Furthermore, it provides the flexibility for assessing in the ESD development for different LoAs and system designs – from a LoA as low as remote control to high as fully autonomous. The flowchart and guidelines can be seen at [5].

Concurrent task analysis

The CoTA developed for the H-SIA method is built over Task Analysis (TA) theory and methods, and expanded to explicitly include the interactions between different parts or agents of the systems. TA was developed in the 1960s [10] and had the initial focus of analyzing human performance. Task analysis is "the collective noun used in the field of ergonomics, which includes HCI, for all the methods of collecting, classifying, and interpreting data on the performance of systems that include at least one person as a system component". Different forms to develop a TA exist, such as Hierarchal TA, Tabular TA, and Cognitive TA [11].

TA allows analyzing complex tasks through the decomposition of goals into sub-goals, so called re-description. The goals and sub-goals are organized in HTA through plans [10]. Plans state the order of the sub-goals to achieve the main goal. From a systems perspective, the HTA should focus on the analysis of the task to understand how the system is supposed to behave and how it may fail. An important element of HTA are the stop rules that determine when to end the re-description. In this work the stop rules are based on the Information Decision Action (IDA) framework.

The IDA model was initially developed as a human behavior model for the operation of nuclear powerplants [12]. It consists of the cognitive phases I (Information collection and pre-processing); D (decision making and situation assessment); and A (action taking). The IDA model has been developed and extended further in recent years [12–16]. It is possible to adapt IDA to different agents of a system. Since the H-SIA method analyzes the interaction between two or more agents, it is beneficial to use a similar model that allows for decomposing functions into the same low-level unit of analysis. In the H-SIA method, thus, IDA model was extended to describe phases and categorize tasks of the autonomous ship as well.

The CoTA consists of several TAs, in which the tasks described as the events in the ESD are re-described until the tasks correspond to one of the IDA phases and the relationship between the sub-task and another agents' task can be established, if this exists. In addition, the CoTA includes a new type of task named "parallel task". Parallel tasks are supporting tasks, i.e., they are necessary for the execution of the other tasks and the interaction between the agents but not explicitly included in the ESD. Parallel tasks are related to the normal operation of the system being executed continuously, not following a specific

order in a plan, i.e., they are executed at the same time with the other tasks. The parallel tasks are normally the ones related to gathering data, monitoring, or communication between the agents.

The CoTA is based on the ESD developed in step 2. The events from the ESD translate into tasks that are performed by the agents. Hence, the ESD presents *what* can happen, and the CoTA further details *how* these events may occur. The CoTA is a success-oriented method that enables the analyst to understand better each agent's tasks that needs to be accomplished for the events of the ESD to take place.

For instance, an event in the ESD may be "Detection of the collision candidate by the autonomous ship". This event is translated into the task "Detect the collision candidate" in the AS' Task Analysis. This task is then re-described using the CoTA stop-rules. The re-description details the sub-level tasks that must be accomplished for the AS to successfully detect the object as a collision candidate, e.g.: gathering and processing data, apply relevant norms, among others.

There are two main approaches when using the CoTA: Analyze the tasks involved in all events of the ESD (i), or to (ii) analyze a specific sequence of events in the ESD scenario. When developed for all the events of the ESD (alternative i), the CoTA provides a detailed overview of how the agents should act to be successful in the possible events of the ESD. The scenario specific CoTA (ii), presents the tasks that should be performed for a success outcome in a specific sequence of events.

The CoTA adopts and expand the HTA plans described in [17]. The CoTA plans describe the order of sub-tasks in order to achieve a successful main task. The CoTA plan may determine for instance a sequence (e.g. 1→2→3 – the tasks 1, 2, and 3 must be performed in this order); or a decision (e.g., Task 1 is performed and, if a condition is satisfied, task 2 is performed; if no, task 3 is performed). In addition, it contains the parallel tasks, and a scenario-specific plan.

The CoTA can be developed from the ESD following the steps below, the relationship between CoTA and ESD is highlighted in Figure 2:

1. Definition of agents to be analyzed, each of the agents will have an HTA;
2. Definition of Task 0, this may be to avoid collision and recover successfully from the initiating event;
3. Definition of agents that are mainly acting in each event agents;

4. Definition of high-level tasks: each event of the ESD translates into a high-level task in each of the respective HTAs. It is recommended to develop a table for correspondence between the event from the ESD and the Task ID in the CoTA;
5. Identification of parallel tasks;
6. Re-description of tasks until stop rules are satisfied. The first rule always must be satisfied, whereas the second may not be satisfied.
   i) The task is associated with only one of the I-D-A phases and, for the dependent tasks;
   ii) The task represents the interaction with another agent.
7. Identification and highlighting of interface tasks.

The CoTA can be used for multiple purposes, such as development of procedures, identification of specific subsystems and components that are necessary for a successful task, identification of failure sources of the human operator or the autonomous system identification of tasks that need to be accomplished for a certain outcome, identification of interface tasks, and analysis of failure propagation.

The scenario specific CoTA

As stated previously, the CoTA may be used for analyzing a specific sequence of events instead of all events of the ESD. This may be achieved from the complete CoTA or directly from the specific sequence of events. In both cases, the development of the scenario specific CoTA starts with the identification of the events involved in the desired ESD path. To make use of the complete CoTA, the analysts identify and selects the tasks of each agent's TA that belong to that sequence. This process may be assisted by the table developed in Step 4. When developing the CoTA from the sequence of events, the analysist follows all the steps outlined above, just for these specific tasks. An example of a scenario specific CoTA is shown in Figure 3.

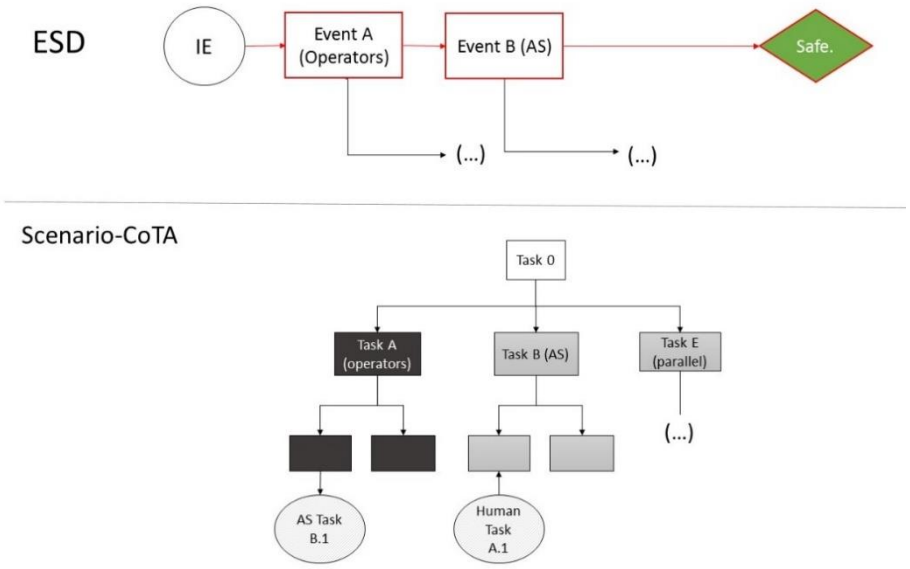


Figure 3: Scenario-specific CoTA example (Adapted from [5])

Discussion and conclusion

In the H-SIA method an autonomous system is analyzed as whole, rather than focusing on each component separately. The process may assist in the comparison of different concepts and designs of an autonomous system. The use of a generic flowchart and generally valid principles produces results that are comparable, reproducible and traceable. An additional benefit of the H-SIA method is the identification and tracking of interdepend tasks of different agents in a system.

The features of the ESD and CoTA makes the H-SIA method a valuable technique for analysis of safety of autonomous systems' operations. It may be used in the design phase, to develop procedures and to derive specifications, for failure events identification, and the results can be further integrated into risk assessments.

Some limitations of the methods are that although the CoTA is developed using clear guidelines and stop-rules, the identification of parallel tasks and the re-description depends also on the analyst. This may lead to different CoTAs when the H-SIA method is used by different analysts. This variability is, in one sense, a limitation of the method. On the other hand, it offers flexibility for the CoTA to be developed and detailed according to the purpose of the analysis.

Future work includes the detailing of the failure events, through e.g., the development of fault trees and BBNs, in a hybrid causal logic model. The method

can benefit from validation through applications to existing autonomous systems and projects, as well as through feedback from experts use.

References

[1] Rødseth ØJ, Tjora A. A risk based approach to the design of unmanned ship control systems. Proceeding Conf Marit Technol 2014:153–62.

[2] Ramos M, Utne IB, Vinnem JE, Mosleh A. Accounting for human failure in autonomous ships operations, Taylor & Francis Group; 2018, p. 355–63.

[3] Ramos M, Utne IB, Mosleh A. Collision avoidance on maritime autonomous surface ships: operators' tasks and human failure events. Saf Sci 2018.

[4] Thieme CA, Utne IB, Haugen S. Assessing ship risk model applicability to Marine Autonomous Surface Ships. Ocean Eng 2018;165:140–54. doi:10.1016/j.oceaneng.2018.07.040.

[5] Ramos M, Thieme CA, Utne IB, Mosleh A. Human-System Concurrent Task Analysis for Maritime Autonomous Surface Ship Operation and Safety. *Submitted to* Reliab Eng &andSystems Saf

[6] Swaminathan S, Smidts C. The Event Sequence Diagram framework for dynamic Probabilistic Risk Assessment. Reliab Eng &andSystems Saf 1999;63:73–90.

[7] Ekanem NJ, Mosleh A, Shen S-H. Phoenix–A model-based Human reliability analysis methodology: Qualitative analysis procedure. Reliab Eng Syst Saf 2015;145:1–15. doi:10.1016/j.ress.2015.07.009.

[8] Ekanem NJ, Mosleh A. Phoenix – A Model-Based Human Reliability Analysis Methodology : Quantitative Analysis Procedure and Data Base. Proc. to Probabilistic Saf. Assess. Manag. PSAM 12, Hawaii: 2014.

[9] Ramos MA. A Methodology for Human Reliability Analysis of Oil Refineries and Petrochemical Plants. Federal University of Pernambuco, 2017.

[10] Shepherd A. Hierarchical Task Analysis. London: Taylor & Francis; 2001.

[11] Annett J, Stanton N. Tasks Analysis. 2000.

[12] Smidts C, Shen SH, Mosleh A. The IDA cognitive model for the analysis of nuclear power plant operator response under accident conditions. Part I: problem solving and decision making model. Reliab Eng Syst Saf 1997;55:51–71. doi:http://dx.doi.org/10.1016/S0951-8320(96)00104-4.

[13] Chang YHJ, Mosleh A. Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents Part 3: IDAC operator response model. Reliab Eng Syst Saf 2007;92:1041–60. doi:10.1016/j.ress.2006.05.013.

[14] Chang YHJ, Mosleh A. Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents . Part 2 : IDAC performance influencing factors model 2007;92:1014–40. doi:10.1016/j.ress.2006.05.010.

[15] Chang YHJ, Mosleh A. Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents Part 5: Dynamic probabilistic simulation of the IDAC model. Reliab Eng Syst Saf 2007;92:1076–

101. doi:10.1016/j.ress.2006.05.012.

[16] Chang YHJ, Mosleh A. Cognitive modeling and dynamic probabilistic simulation of operating crew response to complex system accidents. Part 4: IDAC causal model of operator problem-solving response. Reliab Eng Syst Saf 2007;92:1061–75. doi:10.1016/j.ress.2006.05.011.

[17] Annett J. Hierarchical Task analysis. In: Diaper D, Stanton NA, editors. Handb. Task Anal. Human-Computer Interact., London: Lawrence Erlbaum Associates; 2008, p. 67–82.

Why use Formal Methods for Autonomous Systems?

Matthew Luckcuck

Department of Computer Science, University of Liverpool, United Kingdom
m.luckcuck@liverpool.ac.uk

Formal Methods are mathematically based techniques for software design and engineering, which allow the description of and reasoning about a system's behaviour. Autonomous systems are inherently safety-critical and increasingly being introduced into everyday settings, so using robust development and verification methods is prudent. We argue that formal methods are an effective tool for the development of autonomous systems. They allow unambiguous description of requirements and systems, and are effective in several phases of the engineering life-cycle. Modern formal methods often include highly automated tool support, even older methods have some automation in their tools. Formal Methods should take their place alongside a variety of other robust software engineering techniques for developing autonomous systems.

Introduction

Autonomous Systems (AS) are often embodied in a robotic system and are increasingly being used (or proposed for use) in situations where they are near or interact (physically or otherwise) with humans. This means that AS are safety-critical systems, where failures can cause harm or death. For AS used in industry, the people at risk are likely to be workers; for systems like autonomous vehicles and domestic assistants, the people at risk will be users and bystanders. The security of AS must also be ensured, both because of the sensitive data they are likely to contain and because a security failure can cause a safety failure.

Formal Methods (FM) are mathematically defined techniques for robustly reasoning about systems. They allow the unambiguous description of rules about a system's data and behaviour, and can be applied to the specification, design, and verification phases of systems engineering. Formal Verification (FV) uses FM to show that a system behaves according to a rule (or a property). Like verification by testing or simulation, FV requires a good description of the system and what properties to check for; without this knowledge the verification is unlikely to be valid.

AS need to be safe, correct, and trustworthy, so the most robust design and verification methods available must be used. FM have been used successfully in various industrial projects [1] and there are many academics uses of them to tackle the inherent challenges of AS [2].

The challenges of engineering AS may be external to the system, such as the need to model the system's operating environment and provide robust

evidence of its safety to a regulator. One way to bridge the gap between design-time models of the environment and the real world is Runtime Verification (RV), where a formal description of how the system should behave is used to monitor the running program, an example of which is described in Sect.3. FM are unambiguous and often exhaustive, so provide robust regulatory evidence and traceability.

Other challenges of engineering AS may be internal to the system, for example they could relate to how the autonomy is implemented or how the system may reconfigure itself. Again, FM's unambiguous and exhaustive nature help us to check these complex systems. Some formal models can be animated, allowing a user to step through each state and make any available choices. This is similar to step-by-step debugging of program code, so formal models can be used as rapid prototypes.

The rest of this paper is organised as follows. Section 2 introduces some popular approaches to using formal methods, Sect.3 describes some current examples of applying FM to AS, Sect.4 discusses opportunities for using FM with AS, and Sect. 5 concludes the paper.

Types of Formal Approaches

This section introduces four approaches to using FM, by which we mean the framework(s) or technique(s) used for verification. This is not an exhaustive description of the available approaches, but they are some of the most popular approaches we found in previous work [2].

Model Checking

Model checking checks if a property holds in every state of a formal specification. It is a flexible approach that can use a variety of formal notations, though the most prevalent is temporal logic. Some model checkers accept timed (e.g. Uppaal[7]) or probabilistic models (e.g. PRISM[8]), and there are *program* model checkers (e.g. Java PathFinder[9]) that operate on the program itself [3]

Model checking has some advantages over other approaches: model checkers are automatic, which makes them relatively easy to use; also, the concept of checking every state in a model to see if a required property holds is relatively intuitive. However, because model checking exhaustively explores a specification one must be careful about the input specification and the chosen properties to avoid state space explosion (where the number of states that the

[7] http://uppaal.org/
[8] http://www.prismmodelchecker.org/
[9] https://github.com/javapathfinder

model checker has to search becomes intractable). State space explosion, and building the specifications themselves, are the two obvious overheads to this approach.

Theorem Proving

Theorem Proving is an approach that can produce formal proofs of the correctness of a software system. A variety of logical systems exist for describing a system, and there are powerful tools to aid the user. Formal proofs can be used to provide robust evidence to regulators for the certification of autonomous systems. They also have the advantage of being able to describe systems with an infinite state.

Theorem proving is effective and powerful, but the learning curve for the approach and its tools may be higher than the others we discuss here. Also, the concept and results may be more difficult to explain to stakeholders without formal methods experience.

Runtime Verification

In Runtime Verification (RV) a component called a monitor consumes events from a system and compares them to a formal model of the expected behaviour. If the system's behaviour differs from that described by the model, then the monitor can log the failure, alert the user, or trigger mitigating behaviour.

On its own, RV cannot guarantee a program's behaviour, but it can sidestep the challenge of verifying a complex system. The formal model used is often simpler than if the entire system has been captured, which reduces overhead in developing the model. Runtime verification also helps bridge the *reality gap* (between a model and the real world) by checking formal properties and assumptions at runtime. Because the monitor is another runtime component of the system, it can add to the resource overhead. Reducing this overhead is one of the aims of *Predictive* Runtime Verification [4], [5]

Formal Synthesis

Formal Synthesis is an approach that automatically derives low-level controllers for AS (often movement plans for autonomous robotic systems) from high-level task specifications. The utility of this approach lies automating the conversion of complex task specifications into controllers, when they cannot be trivially converted to a sequence of "go here" statements.

Formal synthesis is an active research area, and can be a powerful technique for deriving a controller that implement a particular task or behave according to certain rules. Various synthesis approaches adapt model-checking

algorithms, but for more complex systems this can also cause state space explosion [6]. Automating controller synthesis moves the development overhead to an earlier phase; like with model checking, developing the right specification is key.

Current Examples

This section describes some recent examples from the literature of applying FM to the inherent challenges of developing AS. This is nowhere near being an exhaustive list, but is intended to show a range of approaches. A fuller account can be found in previous work [2].

Webster et al. [7], [8] tackle an AS controlling a pilotless aircraft. They use a program model checker to check that the AS meets the requirements described in the Rules of the Air, from the UK's Civil Aviation Authority. The Rules of the Air tell pilots how they should fly, so this process replicates licencing a human pilot in the verification of the AS. These studies show how FM can validate natural language requirements (the Rules of the Air) and verify program code.

Ethical concerns must be dealt with if AS are to behave (and be considered) trustworthy. Enabling robots to deal with ethical choices is an open challenge. This is often approached by looking at the choices of an AS and deciding if they are ethical or not [9]–[11]. While this is a useful place to start, real-life situations are unlikely to be so simple. One study that does tackle less dichotomous situations provides a language that captures the ethical weighting of the AS's choices. [12] This language allows the AS to reason about its choices and is amenable to program model checking. This allows the verification of more complex ethical properties, such as in situations where there are only unethical choices and the AS much choose the least bad option.

As mentioned in Sect. 2, RV can help to bridge the reality gap by checking a running program. This has been used to check assumptions made during the system's design using runtime monitors [13]. This enables the authors to validate the formal modelling of the system and the environment it will operate in. RV can also highlight when design-time assumptions about a system's environment become invalid during execution [14].

Finally, several studies have shown the utility of formally synthesising (usually movement) plans that satisfy some specified properties (e.g [15]). Some have tackled generating movement plans for autonomous robots [16], [17]. This approach has the potential to be used during execution by the AS itself [16]. Since this is a similar technique to model checking, it too can suffer from state explosion. Both [16] and [17] plan a short distance ahead to help sidestep this

issue. Formal synthesis has also been applied to systems made of several autonomous robots [6], [18].

Future Directions

This section discusses some potential future uses of FM in the engineering of AS. Generally, AS are deployed to perform a task that a human is currently doing. This presents challenges for verification, because the software is responsible for making more choices; but it also enables new opportunities, like being about to look inside the 'brain' of the system and ensure it's making correct, ethical, and safe decisions. This requires transparent and explainable AS. An example is the concept of an *ethical black box* [19], which records the sensor input and internal state of a system to enable offline checking of the system's decisions in the event of a failure.

As previously mentioned, FM are useful for clarifying what a system should and should not do because they allow the unambiguous description of the system and its requirements. Obviously, this can improve the implementation of those requirements, because they are not open to interpretation like natural-language requirements might be. Additionally, FM are useful as an intermediate language between regulatory frameworks or requirements and a system's design. Again, they will clarify what the system should and should not do, but they also provide traceability of those requirements into the final system. The work of formalising regulatory requirements is not to be underestimated, but once done they would be reusable in building other systems implementing these regulatory requirements.

Formal specifications can also be reused on the same system. We should be considering using some kind of RV in AS to ensure that the system is still fulfilling its original specification. This could be online, or a set of checks rerun at regular intervals. It would be especially useful if the system learns online, is capable of reconfiguring itself, or will be in use autonomously for a long time.

Autonomous Vehicles are a good example of where both formal specification and RV can be used. We have had many decades of learning how to safely develop vehicles, so this is of less concern when developing a vehicle that is autonomous; the challenge lies in developing the AS controlling the vehicle. When a person learns to drive a car, they must (usually) pass some tests to obtain a licence – these are also the requirements of an AS for controlling a car. A formalisation of these requirements gives us what we need to check before giving an autonomous vehicle a 'licence'. More interestingly, checking these requirements regularly allows us to revoke the 'licence' if the checks fail.

Conclusion

This position paper argues that Formal Methods can help engineer safe, correct, and trustworthy Autonomous Systems. It describes a snapshot of what kinds of formal methods exist and some examples of current applications of Formal Methods to different challenges of Autonomous Systems. Finally, it discusses potential uses of Formal Methods in the future.

The discussion here has only talked about applying a single Formal Method to a system, but has hinted at there being other methods (both formal and non-formal) involved in the engineering of one system. In previous work [20] we argue that Autonomous Systems (specifically in the context of autonomous robotic systems) require *integrated* Formal Methods (iFM), which refers to the integration of multiple Formal Methods or of formal and non-formal methods. Joining Formal Methods in this way can produce a formalism able to capture, for example, both static and dynamic behaviour. Blending formal and non-formal methods can produce easy to use approaches, such as graphical notations the are automatically formalised by a tool. In both cases, iFM can allow the combination of the best methods for examining a particular domain.

As previously mentioned, Formal Verification needs good descriptions of systems and their requirements, just like other verification techniques. This means that the Formal Methods and Autonomous Systems communities need to be better at collaborating. Formal Methods need to be able to cope with the architectures and requirements of Autonomous Systems. In turn, Autonomous Systems engineering must ensure that it enables robust verification techniques and works on clear definitions of the system's scope and requirements.

Formal Methods are, by no means, a panacea for the challenges of engineering safe, correct, and trustworthy Autonomous Systems; however, they can provide robust specification and verification of various parts of an Autonomous Systems. Because of the safety and security implications of failures of an Autonomous Systems, the most robust methods must be chosen for each system component and development phase, so Formal Methods should be included as part of the toolbox for engineering Autonomous Systems.

References

[1] J. Woodcock, P. G. Larsen, J. Bicarregui, and J. Fitzgerald, "Formal methods: Practice and Experience," ACM Comput. Surv., vol. 41, no. 4, 2009.

[2] M. Luckcuck, M. Farrell, L. Dennis, C. Dixon, and M. Fisher, "Formal Specification and Verification of Autonomous Robotic Systems: A Survey," 2018.

[3] W. Visser, K. Havelund, G. Brat, S. J. Park, and F. Lerda, "Model Checking Programs," Autom. Softw. Eng., vol. 10, no. 2, 2002.

[4] X. Zhang, M. Leucker, and W. Dong, "Runtime verification with predictive semantics," in Lect. Notes comput. Sci. 2012.

[5] S. Pinisetty, T. Jéron, S. Tripakis, Y. Falcone, H. Marchand, and V. Preoteasa, "Predictive runtime verification of timed properties," J. Syst. Softw., vol. 132, 2017.

[6] Y. Chen, X. C. Ding, A. Stefanescu, and C. Belta, "A Formal Approach to Deployment of Robotic Teams in an Urban-Like Environment," in Distrib. Auton. Robot. Syst., vol. STAR 83, A. Martinoli, F. Mondada, N. Correll, G. Mermoud, M. Egerstedt, M. A. Hsieh, L. E. Parker, and K. Støy, Eds. Springer, 2013.

[7] M. Webster, M. Fisher, N. Cameron, and M. Jump, "Formal Methods for the Certification of Autonomous Unmanned Aircraft Systems," in Lect. Notes comput. Sci., 2011, vol. 6894.

[8] M. Webster, N. Cameron, M. Fisher, and M. Jump, "Generating certification evidence for autonomous unmanned aircraft using model checking and simulation," J. Aerosp. Inf. Syst., vol. 11, no. 5, 2014.

[9] R. Arkin, P. Ulam, and B. Duncan, "An Ethical Governor for Constraining Lethal Action in an Autonomous System," GEORGIA INST OF TECH ATLANTA MOBILE ROBOT LAB, GEORGIA, Technical Report GIT-GVU-09-02, 2009.

[10] R. Arkin, P. Ulam, and A. Wagner, "Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception," Proc. IEEE, vol. 100, no. 3, 2012.

[11] P. Brutzman Donald et al., "Run-time Ethics Checking for Autonomous Unmanned Vehicles: Developing a Practical Approach, Paper and Slideset," in Int. Symp. Unmanned untethered submers. Technol., 2013.

[12] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal verification of ethical choices in autonomous systems," Rob. Auton. Syst., vol. 77, 2016.

[13] A. Aniculaesei, D. Arnsberger, F. Howar, and A. Rausch, "Towards the Verification of Safety-critical Autonomous Systems in Dynamic Environments," Electron. Proc. Theor. Comput. Sci., vol. 232, Dec. 2016.

[14] A. Ferrando, L. A. Dennis, D. Ancona, M. Fisher, and V. Mascardi, "Recognising Assumption Violations in Autonomous Systems Verification," in Int. Conf. Auton. Agents multiagent syst., 2018.

[15] S. Loizou and K. Kyriakopoulos, "Automatic synthesis of multi-agent motion tasks based on LTL specifications," in IEEE conf. Decis. Control, 2004, Vol.1.

[16] S. Karaman and E. Frazzoli, "Sampling-based motion planning with deterministic μ-calculus specifications," in Conf. Decis. Control, 2009, p. 8.

[17] H. Kress-Gazit, T. Wongpiromsarn, and U. Topcu, "Correct, Reactive, High-Level Robot Control," Robot. Autom. Mag., vol. 18, no. 3, Sep. 2011.

[18] Y. Chen, X. C. Ding, A. Stefanescu, and C. Belta, "Formal approach to the deployment of distributed robotic teams," Trans. Robot., vol. 28, no. 1, 2012.

[19] A. Winfield and M. Jirotka, "The case for an ethical black box," in Lect. Notes comput. Sci., 2017, vol. 10454 LNAI,.

[20] M. Farrell, M. Luckcuck, and M. Fisher, “Robotics and Integrated Formal Methods: Necessity Meets Opportunity,” in Integr. Form. Methods, 2018, vol. 11023.

Safety Case and DevOps Approach for Autonomous Cars and Ships

Thor Myklebust[1(*)], Tor Stålhane[2], Geir K. Hanssen[1]

[1] SINTEF Digital, Norway,
[2] Norwegian University of Science and Technology -NTNU, Norway
[(*)] thor.myklebust@sintef.no

During the last years, there has been an increasing use of agile methods when developing safety-critical systems, such as autonomous cars and ships. In the near future, we do also expect that DevOps, which unifies software development (Dev) and software operation (Ops), will be part of this rapidly growing industries. New technology has made it simpler to monitor the operation of safety systems, together with over-the-air updates and upgrades, making DevOps more relevant. DevOps is also a considerable trend for non-critical systems with a growing know-how and tools. However, DevOps, with its frequent changes, make systems' maintainability – e.g., change impact analysis – a more challenging topic than it was earlier.

ISO has developed a functional safety standard for the automotive industry, ISO 26262:2018 while there does not exist a similar standard for ships. In addition, ISO have developed a SOTIF specification for the automotive domain ISO/PAS 21448:2019. Some of the products and systems developed for ships are certified according to IEC 61508. As an example, IEC 61508 is mentioned in MSC.1/Circ.1512:2015 "Guideline on software quality assurance and human-centred design for E-navigation".

Satisfying IEC 26262:2018 for automotive and IEC 61508:2010 for generic systems and components are not sufficient when developing autonomous cars and ships. The reason for this is among other things that they do not include requirements related to important topics like machine learning and deployment over the air, which are technologies that have gained considerable interest over the past years.

A safety case as evidence is required by ISO 26262:2018 and it is suggested to include requirements for a safety case in the next edition of IEC 61508:2010. A safety case could be the main source of evidence, issued by the manufacturer. The safety case includes information related to what has been done both according to safety standards but also have a wider scope than the safety standards as the current editions of the safety standards does not include all relevant requirements for autonomous systems. And equally important, the safety case

includes information regarding the intended use and operation domain together with the limitations of the safety case.

The presentation will include an evaluation of the two main safety software standards ISO 26262-6:2018 and IEC 61508-3:2010 and their weaknesses and lack of guidance related to the safety of autonomous cars and ships.

It has also become more important to move towards a process with more frequent modifications of the safety software, after the cars or ships have been developed, due to e.g., improved operational feedback, technology improvements and security issues, including safe patching. This change is enabled by the flexibility that software-based solutions can offer, and hence, security issues should also be part of the safety case. This is partly included in the railway standard EN 50129:2003 edition and strengthened in the EN 50129:2018 edition. Security was also strengthened in the second edition of ISO 26262:2018 series.

The Agile Safety Case Approach, that are based on both an agile approach and EN 50129:2003, can be an enabler for future DevOps processes.

The Future of Risk in the Context of Autonomous Ship Operation

Nikolaos P. Ventikos(*), Konstantinos Louzis

Maritime Risk Group, Laboratory for Maritime Transport, School of Naval Architecture & Marine Engineering, National Technical University of Athens, Greece
(*) niven@deslab.ntua.gr

One of the main arguments that support the development of autonomous ship operation is the expected improvement of maritime safety by reducing human error and exposing less people to hazardous environments. However, in most likely scenarios, humans will remain in the loop even from a distance, which indicates that, rather than being reduced risk, will simply migrate to another part of the system. Such effects cannot be captured adequately by the traditional risk theoretical and methodological viewpoints that are based on reductionism. The position we take with this paper is that risk analysis needs to embrace a systemic perspective that accounts for the complexity of autonomous ships and related uncertainties. We support this, first by considering why autonomous ships should be considered as complex systems and secondly by discussing the application of systems-based risk perspectives and relevant methodological frameworks in the autonomous ship domain. Our main conclusion is that future research efforts should provide context-specific systemic models at various levels of abstraction to capture the wide-ranging effects of autonomous ship operation.

Keywords: Maritime safety, Risk Analysis, Autonomous ships, Systemic risk, Uncertainty

Introduction

Several actors in the maritime industry are currently exploring the possibilities for operating ships with various levels of autonomy. Although these concepts are seen by many as too ambitious to be implemented at full scale, the main argument for autonomous operation is that it will improve maritime safety [1] by reducing human error, while exposing fewer people to hazardous environments [2]. The projected benefits for safety are based on the evident apportioning of human error as the main root cause on more than 80% of all marine accidents [3].

This argument is based on a rather simplistic and narrow perspective that equates the absence of people onboard with the absence of safety risk. However, as unmanned and unmonitored autonomous ships are not likely to become an imminent reality, humans will remain in the loop; e.g. relocated to a shore-based facility where they will monitor and intervene if necessary [4–6]. In addition, Wróbel et al. [7] have drawn attention to the fact that even though autonomous ships may reduce navigational accidents (i.e., collisions and groundings) the consequences of other types of accidents, such as fires, may be more extreme without the mitigating action of the human element onboard. Considering that there is currently no firm evidence to prove that autonomous ships will operate

at an acceptable level of safety when deployed in full scale, reliable and well justified risk analysis may provide the industry with the decision support it needs to further justify the development of autonomous ship operation. However, the risk analysis domain has experienced an asynchronous development between thinking and practice. Established thinking leans towards the systemic perspective, acknowledging complex component interactions, while practice has lagged behind with tools that depend on sequential and linear models [8].

The position we take with this paper is that risk analysis in the context of autonomous ship operation needs to consider the complex nature of these systems by employing systemic models and providing informative statements on related uncertainties. The rest of this paper is structured as follows. Section 2 discusses the features of autonomous ships that make them complex in relation to formal definitions of complexity. Section 3 discusses how complexity impacts risk analysis by presenting selected alternative systems-based risk perspectives and reviewing the available frameworks for modelling risk in complex systems. The paper concludes with some comments on how risk analysis may be applied in the context of autonomous ships in a way that addresses their complexity and will provide useful results.

Complexity in the context of autonomous ship operation

Complex systems exhibit certain properties that include non-linearity, the presence of feedback loops that affect their behaviour, self-organization, robustness, emergence, hierarchical organisation, and numerosity of components [9]. In the context of Normal Accident Theory (NAT), Perrow [10] has defined a complex system as one that includes complex interactions and tight couplings. In a risk assessment context, Johansen and Rausand [11] have noted that the common basis among the different definitions of complexity relates to a limited ability to understand and predict complex system behaviour just by understanding the behaviour of its components.

Autonomous ships will have at least some of these properties and therefore we may characterize them as complex systems. The most obvious is numerosity [12], as autonomous ship operation will be enabled by the collaboration between complicated (and potentially complex in themselves) shipboard systems and shore-based systems. Numerous shipboard systems will replace the functionalities of onboard crew, such as providing situational awareness (external and internal), evaluating alternatives for collision avoidance, communicating with the shore, and actuating commands. A shore-

based facility will most likely be used for monitoring and control that will also consist of liveware (a term that refers to the human element that was originally used in computer science), hardware, and software. In this sense, autonomous ship operation might very well be distributed in space with a strong functional dependence between sea and shore, whose malfunctioning or disruption may lead to a system-level failure.

Other properties of complex systems, such as feedback loops and self-organization, emerge more clearly when considering autonomous ship operation within the Maritime Transportation System (MTS). For example, if we consider a possible MTS that includes conventional ships and ships with various levels of autonomy up to artificially intelligent ships, then the interactions between them, independent of central control and subject to multiple feedback loops, may possibly lead to the emergence of different high-level behaviours. From this perspective, non-linearity emerges as ensuring the safety of individual ships does not necessarily ensure the safety of the wider system.

Risk analysis in a complex setting

Complexity impacts how we perceive and analyse risk. Jensen and Aven [13] note that for complex systems we cannot necessarily improve our knowledge at the system level by understanding the individual components, which implies that for analysing risk in a complex system we must have some level of knowledge on the complex interactions. By ignoring interaction, we risk being surprised by accidents that involve independently non-failing components [14]. For example, if we analyse the risk of technical failures and human error separately, which is the approach that has been largely followed in the practice of risk analysis, then we explicitly disregard the effect of their interaction on the risk level of the system through risk aggregation. Furthermore, understanding complex interactions may help to identify how risk migrates from one part of the system to another, such as in the case of transferring control from the ship to a remote centre of operations.

Drawing on the framework of systems engineering, Haimes [15] has provided a definition of risk in complex systems as a vector with the same units as the consequences that is a function of time, the probability of initiating events, the probability of the consequences conditioned upon the initiating event, the vector of the states of the system, and the vector of the consequences. Another interesting definition formulated by Andretta [16], shown in Equation 1, links risk (R) for targets of interest (T_i) to the probability of an anomalous state of the system (ST_a), which results in a damage of specific magnitude (M_d), which in turn results in an adverse effect (E_a). This definition is similar to the one formulated

by the risk triplet of Kaplan and Garrick [17], but the risk scenario is described as a function of anomalous system states that may produce damages that affect targets of interest with specific consequences.

$$R_{\{Ti\}} = P_{\{Ti\}}(ST_a, M_d, E_a) \quad (1)$$

These two definitions of risk are similar in the way they condition risk quantification upon defining a model of the system and determining its states that may have an adverse impact on something of value. Although these definitions provide a solid conceptual framework for characterising risk in complex systems, they do not offer explicit instruction as to how to model the system and the interactions among the components.

Frameworks that address the modelling problem from a systemic perspective include Rasmussen's Risk Management Framework [18], the Systems-Theoretic Accident Model (STAMP) by Leveson [14], the Functional Resonance Analysis Method (FRAM) by Hollnagel [19], and the Event Analysis of Systemic Teamwork (EAST) by Stanton et al. [20]. Strictly speaking, Rasmussen and Leveson provide models for describing how accidents in complex systems occur, while Hollnagel and Stanton provide a way to structure model representations of how the system works. Both Rasmussen and Hollnagel adopt a functional (instead of a structural) decomposition of the system and acknowledge that performance varies. Leveson addresses safety modelling as a control problem and therefore uses a control structure with feedback loops to represent the system. Stanton et al. take a different approach by modelling the system as a network that consists of three sub-networks (social, task, and information) that represent "distributed cognition".

These systemic methodologies currently occupy a very small space in the risk assessment literature [8] and, similarly, in the maritime risk domain there are only a few examples of such applications. In the domain of autonomous ships, Wróbel et al.[21] applied the System Theoretic Process Analysis (STPA) methodology, which is derived from the STAMP framework, on a generic remotely-controlled merchant vessel. The methodology they applied involved modelling the system as a safety control structure, based on information from the available literature and from brainstorming sessions with industry experts. In addition, the authors conducted an uncertainty analysis for the structure of the model by applying the framework proposed by Flage and Aven [22].

Although the systemic perspective on risk is beneficial for the analysis of complex systems, its practical usefulness is not without limitations. This approach has a strong dependence on how to model the system, which has been

criticized by Aven [23] by supporting that the basic definition of risk should not be so strongly conditioned on modelling. Another implication of this dependence is that it may result in inconsistent risk analyses as different modelling approaches may include different component interactions. In addition, in cases of novel and innovative systems, such as autonomous ships, the limited knowledge of their exact structure adds epistemic uncertainties into the analysis, which should be explicitly identified. An important issue with the systems-based modelling frameworks is that they do not explicitly support the quantification of risk and are therefore practically limited in hazard identification, with the exception of EAST that includes Social Network Analysis (SNA) metrics [24]. In fact, they are sometimes purposefully non-quantifiable, see for example Leveson [14], expressing an abandonment of the probability concept in light of its limitations especially. In a critical view of this issue, in the cases of STAMP and FRAM, Bjerga et al. [25] have pointed out that these approaches do not handle uncertainty adequately and support that some expression of uncertainty (be it probabilistic or of some other form) over the future manifestation of the risk scenarios should be included. Risk quantification is however inevitably useful for supporting decisions, and if not explicitly specified it is left up to the decision maker to judge the likelihood of the risk scenarios.

Conclusions

The argument for improving maritime safety with autonomous is largely based on a simplistic assumption that equates the absence of people onboard with the absence of risk. However, this assumption may not be true considering the complexity of autonomous ships and the increased complexity of the MTS due to their operation. Complex interactions and strong dependencies among risk factors will most likely produce emergent behaviours that may impact safety in ways that are difficult to predict. Risk analysis for autonomous ships should therefore account for this complexity from both a risk theoretical and methodological viewpoint.

The systemic risk perspectives that have been reviewed provide a strong link between the risk of unwanted consequences and a model of the system based on the description of system states. This perspective may be combined with the existing methodological frameworks that provide guidance on how to structure a system model. These frameworks provide a solid basis for modelling but need to be supplemented by appropriate statements of uncertainty. The application of such approaches in the domain of autonomous ships may prove useful, which is a position that has been acknowledged in the relevant literature. However, there is a need to create more context-specific models that capture the important

factors that differentiate autonomous from conventional ship operation. Finally, modelling attempts need to be made at various levels of abstraction (from individual ships to the MTS) to determine the potential wide-ranging effects on maritime safety from the operation of autonomous ships.

References

[1] Rødseth ØJ, Burmeister H-C. Developments toward the unmanned ship. Proc. Int. Symp. Inf. Ships—ISIS 2012, Hamburg: 2012.

[2] Burmeister H, Bruhn WC, Rødseth ØJ, Porathe T. Can unmanned ships improve navigational safety ? 2014.

[3] Allianz Global Corporate & Specialty. Global Claims Review - Liability in Focus (Loss trends and emerging risks for businesses) 2017.

[4] Ahvenjärvi S. The Human Element and Autonomous Ships. TransNav, Int J Mar Navig Saf Sea Transp 2016;10:517–21. doi:10.12716/1001.10.03.18.

[5] Ramos M, Utne IB, Vinnem JE, Mosleh A. Accounting for human failure in autonomous ships operations, Taylor & Francis Group; 2018, p. 355–63.

[6] Relling T, Lützhöft M, Ostnes R, Hildre HP. A human perspective on maritime autonomy. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10916 LNAI, Springer Verlag; 2018, p. 350–62. doi:10.1007/978-3-319-91467-1_27.

[7] Wróbel K, Montewka J, Kujala P. Towards the assessment of potential impact of unmanned vessels on maritime transportation safety. Reliab Eng Syst Saf 2017. doi:10.1016/j.ress.2017.03.029.

[8] Dallat C, Salmon PM, Goode N. Risky systems versus risky people: To what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. Saf Sci 2016. doi:10.1016/j.ssci.2017.03.012.

[9] Ladyman J, Lambert J, Wiesner K. What is a complex system? Eur J Philos Sci 2013. doi:10.1007/s13194-012-0056-8.

[10] Perrow C. Normal accidents: Living with high risk technologies. 1999. doi:10.5465/AMR.1985.4278477.

[11] Johansen IL, Rausand M. Defining complexity for risk assessment of sociotechnical systems: A conceptual framework. Proc Inst Mech Eng Part O J Risk Reliab 2014. doi:10.1177/1748006X13517378.

[12] Utne IB, Sørensen AJ, Schjølberg I. Risk Management of Autonomous Marine Systems and Operations. Proc 36th Int Conf Ocean Offshore Arct Eng 2017:1–10.

[13] Jensen A, Aven T. A new definition of complexity in a risk analysis setting. Reliab Eng Syst Saf 2018. doi:10.1016/j.ress.2017.11.018.

[14] Leveson N. Engineering a safer world: systems thinking applied to safety. The MIT Press; 2016. doi:10.5860/choice.49-6305.

[15] Haimes YY. On the complex definition of risk: A systems-based approach. Risk Anal 2009. doi:10.1111/j.1539-6924.2009.01310.x.

[16] Andretta M. Some Considerations on the Definition of Risk Based on Concepts of Systems Theory and Probability. Risk Anal 2014. doi:10.1111/risa.12092.

[17] Kaplan S, Garrick BJ. On The Quantitative Definition of Risk. Risk Anal 1981. doi:10.1111/j.1539-6924.1981.tb01350.x.

[18] Rasmussen J. Risk management in a dynamic society: A modelling problem. Saf Sci 1997. doi:10.1016/S0925-7535(97)00052-0.

[19] Hollnagel E. FRAM: The Functional Resonance Analysis Method. CRC Press; 2017. doi:10.1201/9781315255071.

[20] Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, Jenkins DP. Human Factors Methods. CRC Press; 2017. doi:10.1201/9781315587394.

[21] Wróbel K, Montewka J, Kujala P. System-theoretic approach to safety of remotely-controlled merchant vessel. Ocean Eng 2018;152:334–45. doi:10.1016/j.oceaneng.2018.01.020.

[22] Flage R, Aven T. Expressing and communicating uncertainty in relation to quantitative risk analysis. ... Risk Anal Theory Appl 2009.

[23] Aven T. On Some Recent Definitions and Analysis Frameworks for Risk, Vulnerability, and Resilience. Risk Anal 2011. doi:10.1111/j.1539-6924.2010.01528.x.

[24] Stanton NA, Harvey C. Beyond human error taxonomies in assessment of risk in sociotechnical systems: a new paradigm with the EAST 'broken-links' approach. Ergonomics 2017. doi:10.1080/00140139.2016.1232841.

[25] Bjerga T, Aven T, Zio E. Uncertainty treatment in risk analysis of complex systems: The cases of STAMP and FRAM. Reliab Eng Syst Saf 2016. doi:10.1016/j.ress.2016.08.004.

A Survey on Autonomous Vehicles Interactions with Human

Bentolhoda Jafary[1(*)], Lance Fiondella[1], Ali Mosleh[2]

[1]University of Massachusetts Dartmouth, Dartmouth, USA;
[2]University of California Los Angeles - UCLA, Los Angeles, USA
[(*)] bjafary@umassd.edu

Autonomous vehicles (AVs) or self-driving cars have the potential to replace human-operated cars. AVs can sense the environment and even navigate some of the roads in conditions humans find challenging. This may quickly lead to people's over reliance on AVs and overconfidence that no failures will occur. Therefore, AVs can impact society positively and negatively. AVs are X-ware systems that consist of software, hardware, humans, and their interactions. Despite the large number of studies on AVs, there are still a large number of unsolved problems. One major challenge for AVs is communication with driver as well as pedestrians. Most of the previous research efforts consider software failures, whereas few consider the role of humans in the current transition to a society in which self-driving cars predominate. This paper considers the interaction between the AVs and humans including: I) the driver and passenger of the AV and II) pedestrians. We also discuss related studies on human behavior.

Keywords: Driver-pedestrian interaction, Human intention, Behavior analysis

Introduction

Technological advances, such as artificial intelligence, are being leveraged to build our future smart cities with the intelligent infrastructure in which driverless vehicles will be the key feature of the transportation network. Commercial cars are categorized into 5 levels [1], including: (i) Level 1 cars which are entirely manual; (ii) Level 2 cars in which only single operations such as anti-lock braking, brake assist, and electronic stability are automated; (iii) In level 3 cars, called combined function automation, two or more functions are automated; (iv) Level 4 cars are those which do not require attention of the driver at any time because they use automation to control all aspects of the driving task for extended periods; (v) Finally, level 5 cars are driverless and completely automatic. Nowadays autonomous vehicles (AVs) or self-driving cars (level 4 and 5 cars) are in the research spotlight in academia and of great interest to giant companies such as Apple, Google, Tesla, Uber, and Volvo [2]. AVs can sense the environment and navigate the roads even in conditions that are challenging for humans to manage.

There have been numerous successes, since the early attempts at autonomy [3] and several studies on autonomous vehicles have been published. Since 2004, the Defense Advanced Research Projects Agency (DARPA) has held three major challenges on robotic vehicles [4]. In 2007, the DARPA urban

challenge focused on the research and development of robot cars for urban environments, which had to navigate moving traffic safely while obeying California traffic regulations. However, they excluded pedestrians and bicyclists [5] in their research. Later, Nothdurft et al. [6] introduced "Leonia" in the Stadtpilot project, an autonomous vehicle, which demonstrated the ability to drive autonomously in real urban conditions. They discussed the legal issues of driving AVs such as the role of driver, safety, and control concepts. Mark et al. [7] reviewed some of the main technologies and architecture of autonomous vehicles, and brought some of the emerging challenges and opportunities into consideration, including navigation system, software integration, and algorithmic integration. Bagloee [8] reviewed the challenges and opportunities that autonomous vehicles might create and, discussed the possible advantages and disadvantages of the AVs. Bimbraw [9] reviewed the basic chronology of autonomous vehicle technology. Tian et al. [10] proposed a tool for automated testing of a Deep-Neural-Network-driven Autonomous Car capable of detecting behavior that could lead to crashes. Panichella et al. [11] proposed a technique to detect the feature interaction failures in the context of autonomous vehicles by developing new search-based test generation algorithm.

Despite the large number of studies on AVs, the research on the interaction between human and AVs is scarce yet indispensable [12]. Driving in an urban area is challenging because there are more pedestrians in this area, which requires special considerations for AVs to be compatible in such an uncertain environment. Moreover, AVs must interact with the other users of the road and human-operated vehicles. Therefore, it is crucial to consider the challenges that driver and AV's passengers, and pedestrians will face. This paper reviews the relevant literature on these areas with special focus on providing a better understanding of the role of human interaction with AVs

The remainder of this paper is organized as follows: Section 2 reviews the literature on the impact of AVs on pedestrians. Section 3 reviews the interaction between AVs and driver. Section 4 provides conclusions and offers directions for future research.

Interaction between AVS and pedestrians

Evidence suggests that autonomous vehicles are more cautious around pedestrians. Google's autonomous vehicles collision reports indicate that in most accidents the vehicles are hit from behind because Google's cars stop to give the right-of-way to the pedestrians [13]. Millard-Ball [14] analyzed the interaction between the pedestrians and autonomous vehicles focusing on yielding at crosswalks using game theory. Autonomous vehicles are programmed to respect the right-of-

way of pedestrians, which is conditional on AVs "playing nice." Hulse et al. [15] surveyed almost 1,000 participants to assess their perceptions on safety and acceptance of AVs. The results indicate that pedestrians believe AVs are less risky compared to human-operated cars. Moreover, gender, age, and risk-taking personality play an important role in AV acceptance. For example, females were less comfortable with AVs than males and young adults.

In the case of level 5 AVs, walking could become more pleasant because on-street parking is anticipated to disappear, since driving will become a service and parking move to the suburbs. Moreover, crossing the street should be more convenient, since the AVs must stop for pedestrians and cannot claim that they did not see a pedestrian or drive under the influence of alcohol. Meeder et al. [16] discussed the impact of AVs on pedestrian activity. They identified the potential positive impacts of AVs on pedestrians. For examples, AVs exhibit a higher success rate in detecting the pedestrians compared to human-operated cars. Therefore, walking could be safer and more attractive for pedestrians, since they could cross the street with greater confidence. Furthermore, car pollution will decrease since most of the AVs are expected to be electric. Therefore, the quality of air will improve, the noise level will decrease, and the environment would become even more pleasant for pedestrians. Moreover, more space is available for the pedestrians since the size of the AVs are smaller and they can drive within narrower lanes. It is also anticipated that car sharing will be widespread.

Other researchers have mentioned the negative impacts of AVs on pedestrians. For example, Meeder et al. [16] discussed potential abuse of AVs by pedestrians who could make them stop at every location, which would increase congestion, and the pedestrians would have to take the longer paths as they would likely be banned from not cutting through the AVs' roadways at every location. More importantly, communication between AVs and pedestrians is different. Thus, pedestrians would need to learn new rules, which they may resist. If AVs are more convenient, their use for short trips may be preferred instead of walking, which will increase congestion and degrade the pedestrian experience. Cities may be more organized, and it is unclear how attractive the city center will be to different types of business. Furthermore, a driver's license may no longer be needed and even children could have their own private car. As a result, the number of autonomous cars may increase rapidly, and walking areas may be dominated by AVs.

In contradiction to those who believe that AVs will be more cautious and accurate around pedestrians, others believe that AVs are more likely to be the cause of a crash. Of course, the debate is ongoing. In this regard, one of the central concerns is that AVs are not able to distinguish between different types of objects they

encounter with sufficient accuracy, which may threaten the life of pedestrians and lead to incidents with serious consequences. Additionally, at this stage of automation and the current conditions of the roads and traffic signs, AVs are susceptible to be adversely affected by pedestrians such that some people such as a gang could simply stand in front of an AV or attack the car, in order to steal it. In this case, security cameras on the cars with the ability to communicate with a police station would be beneficial. Moreover, the physical design of an urban area needs to be remodeled, in order to control the interaction between pedestrians and AVs to some extent, which may increase the complexity of street design and create subsequent problems. In such case, individuals will have to learn the new traffic signs and rules that requires time and impact transportation safety.

Interaction between AVS and driver

One of the critical deficiencies of AVs is the driver's susceptibility to erroneous or delayed decision-making following an alarm from the car, similar to those observed in the aviation industry. In order to avoid undesirable consequences, this process of decision-making and taking the appropriate action should usually take place in a very short time window after the alarm goes off. In addition, the frequency of these incidents likely to be orders of magnitude greater than the aviation industry, given the number of cars on the road relative to the number of planes in the sky. This process of recovering after AV's inability to continue to operate can be impaired by a myriad of factors such as the driver being distracted (taking a nap, talking with other passengers, reading, etc.,) or being unconscious of type of failure such as a malfunction in the speed control system.

In automated driving, the driver may be deeply engaged in other activities, thus bringing a distracted driver back into the control loop can become very challenging. In fact, transitions between the human and automated driving is a key design issue for autonomous vehicles. Merat et al. [17] employed a driving simulator to investigate the ability of drivers to handle conditions where automation reverts to manual control, which was based on the length of the time the driver was not looking at the road ahead. They considered eye movement patterns and showed that drivers exhibited the best performance when the control transferred after six seconds after a take-over request. Moreover, they discussed the importance of designing effective human machine interfaces in automated driving conditions, which certify the time and manner in which the message regarding transfer to the manual control is issued. Another imperative factor is how to warn the driver, for example, the necessity of clear language [18] to unambiguously communicate the level of urgency to the driver. Politis et al. [19] considered a language-based warning model to switch from

autonomous to manual control. They evaluated the audio, tactile, and visual warnings and concluded that it is critical for the driver to intuitively understand the level of urgency.

From a human factors perspective, the crucial challenges are designing automation in a way that drivers fully understand the functionalities, capabilities and limitations of the vehicle, and how to keep the driver engaged to maintain situational awareness of what the vehicle is doing and when manual intervention is needed. Cunningham and Regan [20] reviewed some of the human factors challenges in this regard including driver inattention and distraction, skill degradation, and motion sickness. Petermeijer et al. [21] reviewed the literature on vibro-tactile displays as a possible method to alert the driver at the time of transition from automated to manual driving. Four dimensions were considered, including frequency, amplitude, location, and timing. Although vibrotactile feedback has benefits, it also has several limitations such as differences in the response threshold of individuals to receive notice and duration or intensity of vibration that may be uncomfortable. Lu et al. [22] proposed a theoretical framework and investigated the human factors in transition from automated to manual driving by defining different joint driving states of driver and vehicle. Kyriakidis et al. [23] interviewed 12 expert researchers in the field of human factors and discussed the role of human factors in AVs. They identified the commonalities and perspectives regarding human factors. It was recommended that drivers be trained to be aware of AV limitations to ensure they are capable of operating AVs and maintain control of the car in case of transition from autonomous to manual driving.

Clark et al. [24] analyzed the impact of level of distraction with respect to the age of drivers when predicting the performance of taking control of a highly automated vehicle. They showed that younger drivers were more easily distracted than older drivers. Moreover, age and speed were negatively correlated with high speed among younger drivers. However, their study had some limitations, such as small sample size and the type of activities that participants were engaged in to achieve different level of distraction, which may have resulted in limitations to the generalizability of results. Vogelpohl et al. [25] studied the behavior of distracted drivers as they reacted to the unexpected traffic events. Their results indicated most participants reacted to the unexpected conditions and deactivated automation after seven to eight seconds. Moreover, drivers of the automated vehicles exhibited a delay, up to five additional seconds before the first gaze into the mirror and road in comparison with the drivers of the manual cars.

Another significant factor that needs to be considered is the driver's level of skill. It is critical that driver be able to respond in case of automation failure. Lack of

driving skills can be serious and may threaten the life of pedestrians, drivers, and passengers of an AV, although some other factors such as gender, age, level of consciousness are also significant.

As discussed, operating an AV will allow the driver to be easily distracted. Therefore, the time to recognize AV failure and resume manual control will increase. One solution is to use eye detection technology that can track the driver's eyes and alert the driver when the driver is not focused enough. Since reaction time plays a critical role in case of automation failure, it would also be valuable if AVs could predict when something might go wrong and alert in advance.

Conclusion

This paper considers two categories of AV interactions including: I) pedestrian and II) drivers and passengers. The recently published papers in this area were reviewed and the gaps requiring additional focus were identified. Most studies assume AVs will play nice. Although this assumption simplifies the experiments, AVs experience failures, which create unforeseen problems. More studies regarding interaction with pedestrians are needed to develop methodologies and algorithms so that AVs can make robust decisions on what action or sequences of actions would mitigate consequences when confronted with challenging situations. Moreover, current transportation networks are not designed for AVs. Therefore, AV interactions with pedestrians have not been considered in the process of their design. Another category is the interaction between an AV and drivers and passengers of that AV. The driving skills and possible loss of situational awareness of the driver need to be studied systematically to increase the reliability of AVs. For example, a driver with low or degraded skills from lack practice, may perform an incorrect action in a situation that could lead to a collision. Moreover, since driving an AV may be a monotonous task, the driver may become easily distracted by other activities making them more prone to taking inappropriate actions when human intervention is required.

Future work will consider the impact of AVs on pedestrians, drivers, infrastructure and other users of the road. More specifically, we will discuss the possible failures in greater detail and will offer potential solution and methods to objectively measure efforts to make improvements that enhance safety and convenience

Acknowledgment

This research was partially supported by the B. John Garrick institute for the Risk Sciences, University of California, Los Angeles.

References

[1] S. Casner, E. Hutchins, and D. Norman. "The challenges of partially automated driving." Communications of the Association for Computing Machinery, 59.5 (2016): 70-77.

[2] L. Hook, and R. Waters. "Google's Waymo passes milestone in driverless car race." Financial Times. https://www.ft.com/content/dc281ed2-c425-11e7-b2bb-322b2cb39656. Accessed 28 (2018).

[3] F. Kroger, "Automated driving in its social, historical and cultural contexts," in Autonomous Driving. New York, NY, USA: Springer-Verlag, 2016, pp. 41–68.

[4] U. Ozguner, C. Stiller, and K. Redmill. "Systems for safety and autonomous behaviour in cars: The DARPA Grand Challenge experience." Proceedings of the IEEE 95.2 (2007): 397-412.

[5] C. Reinholtz et. al. DARPA: Urban Challenge Technical Evaluation Criteria. Technical report, DARPA, Arlington, VA, USA (2006)

[6] T. Nothdurft, P. Hecker, S. Ohl, F. Saust, M. Maurer, A. Reschka, and J. Rudiger Bohmer. "Stadtpilot: First fully autonomous test drives in urban traffic." International IEEE Conference on Intelligent Transportation Systems, 2011.

[7] M. Campbell, M. Egerstedt, J. How, R. Murray. "Autonomous driving in urban environments: approaches, lessons and challenges." Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 368.1928 (2010): 4649-4672.

[8] S. Bagloee, M. Tavana, M. Asadi, and T. Oliver. "Autonomous vehicles: challenges, opportunities, and future implications for transportation policies." Journal of Modern Transportation 24.4 (2016): 284-303.

[9] K. Bimbraw. "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology." International Conference on Informatics in Control, Automation and Robotics (ICINCO), Vol. 1. IEEE, 2015.

[10] Y. Tian, K. Pei, S. Jana, and B. Ray. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars." Proceedings of the International Conference on Software Engineering. ACM, 2018.

[11] A. Panichella, S. Nejati, L. Briand, T. Stifter, et al. "Testing Autonomous Cars for Feature Interaction Failures using Many-Objective Search." Proceedings of the 33rd IEEE/ACM International Conference on Automated Software Engineering, 2018.

[12] I. Wolf. "The interaction between humans and autonomous agents." Autonomous

driving. Springer, Berlin, Heidelberg, 2016. 103-124.

[13] E. Aria, J. Olstam, and S. Christoph. "Investigation of automated vehicle effects on driver's behavior and traffic performance." Transportation Research Procedia 15 (2016): 761-770.

[14] A. Millard-Ball. "Pedestrians, autonomous vehicles, and cities." Journal of planning education and research 38.1 (2018): 6-12.

[15] L. Hulse, H. Xie, and E. Galea. "Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age." Safety Science 102 (2018): 1-13.

[16] M. Meeder, E. Bosina, and U. Weidmann. "Autonomous vehicles: Pedestrian heaven or pedestrian hell?" Swiss Transport Research Conference, 2017.

[17] N. Merat, A. Jamson, F. Lai, M. Daly, and O. Carsten. "Transition to manual: Driver behaviour when resuming control from a highly automated vehicle." Transportation research part F: traffic psychology and behaviour 27 (2014): 274-282.

[18] C. Baldwin, and M. Colleen. "Perceived urgency, alerting effectiveness and annoyance of verbal collision avoidance system messages." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 46. No. 22. Sage CA: Los Angeles, CA: SAGE Publications, 2002.

[19] I. Politis, S. Brewster, and F. Pollick. "Language-based multimodal displays for the handover of control in autonomous cars." Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2015.

[20] M. Cunningham, M. A. Regan. "Autonomous vehicles: human factors issues and future research." Proceedings of the Australasian Road Safety Conference. 2015.

[21] S. M. Petermeijer, J. CF De Winter, and K. J. Bengler. "Vibrotactile displays: A survey with a view on highly automated driving." IEEE Transactions on Intelligent Transportation Systems 17.4 (2016): 897-907.

[22] Z. Lu, R. Happee, C. Cabrall, M. Kyriakidis, and J. de Winter. "Human factors of transitions in automated driving: A general framework and literature survey." Transportation research part F: traffic psychology and behaviour 43 (2016): 183-198.

[23] M. Kyriakidis, J. de Winter, N. Stanton, T. Bellet, B. van Arem, K. Brookhuis, M. Martens, K. Bengler, J. Andersson, and N. Merat et al. "A human factors perspective on automated driving." Theoretical Issues in Ergonomics Science (2017): 1-27.

[24] H. Clark, A. McLaughlin, B. Williams, and J. Feng. "Performance in takeover and characteristics of non-driving related tasks during highly automated driving in younger and older drivers." Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 61. No. 1. Sage CA: Los Angeles, CA: SAGE Publications, 2017.

[25] T. Vogelpohl, M. Kühn, T. Hummel, T. Gehlert, and M. Vollrath. "Transitioning to manual driving requires additional time after automation deactivation." Transportation research part F: traffic psychology and behaviour 55 (2018): 464-482.

The Management of Risk in Autonomous Marine Ecosystems – Preliminary Ideas

Sunil Basnet(*), Osiris A. Valdez Banda, and Spyros Hirdaris

[1] Marine Technology, Department of Mechanical Engineering, Aalto University, Finland.
(*) sunil.basnet@aalto.fi

Marine industry is set to experience a change of an era as the development of autonomous ships has already started. However, the operation of autonomous ships is not possible unless the new types of hazards and its associated risks due to the rapid technological changes are identified and controlled. Thus, it is necessary to identify or develop a suitable risk management model that can identify these new types of risks.

This paper aims to identify and suggest a suitable risk management model or a category of models for managing the risks in autonomous marine ecosystems. Firstly, the available models and their categories in all major domains such as aviation, automotive, railway and marine industry are explored. Then, a SWOT analysis is conducted for each model category to assess the strengths, weaknesses, opportunities and threats. The results of the SWOT analysis show that the systemic models such as STAMP can be a suitable option than the traditional categories such as sequential and epidemiological models.

Introduction

Because of the continuous development of autonomous technologies, the marine industry currently explores feasible options for the design and operation of maritime autonomous systems [1]. Cross-modal technology disruption trends imply new risks. Hence, it becomes essential to understand gaps in existing risk management systems and explore the potential of novel risk assessment methods especially considering societal and industry expectations for sustainable life cycle solutions [2].

Whereas the marine industry traditionally utilized operational data to understand risks, autonomous marine systems are new and their development is based on limited databases. A direct influence of this is that quantitative risk assessment (QRA) methods and passive risk management practices become less relevant [3]. This is the reason why there is a need to develop new dynamic risk assessment models that are suitable for detecting multiplicity of risks implied by

the impact of disruptive technologies, limited human – machine interaction and limited in service experience.

At first instance, the limited availability of data and experience in the maritime domain suggest the opportunity to learn from other industries such as automotive, railway and aviation. As a first step toward this direction, this paper aims to explore the potential of available cross-modal risk management methods and frameworks and then suggest some initial thinking directions in terms of developing techniques and models that may be more suitable for the marine domain.

Methodology

In this paper the available hazard and accident analysis models for risk management used by aviation, railway, automotive and maritime domains are explored and models are then classified based on the taxonomy suggested by Underwood and Waterson [4]. A SWOT analysis is then performed for each category with the aim to understand their potential of implementation. The details of the SWOT analysis and the detailed review of other transport domains are presented in Manzur et al. [5].

Literature Review – Exploring hazard and accident analysis models in major domains

Over the years various systems analysis models and tools have been developed. Figure 1 presents the timeline of the best-known risk management methods [6]. From a critical review perspective, it appears that the railway industry has been leading the way in terms of implementation. For example, highly - automated systems such as the magnetic track inspection systems have been introduced since 1910 with the aim to supplement human inspection [7]. Railway regulatory bodies have recommended the usage of traditional methods such as Event Tree Analysis (ETA), Fault Tree Analysis (FTA), Failure Mode and Effect Analysis (FMEA) and Hazard and Operability study (HAZOP) for managing the risks of modern trains with the higher implementation of automated systems. Recently, Belmonte et al. [8] and Dong [9], suggested the implementation of modern methods such as the Functional Resonance Accident Method (FRAM) and System-Theoretic Accident Model and Processes (STAMP). It is believed that these modern methods may present a good addition to the classical approaches as they cover the complex interactions of modern systems.

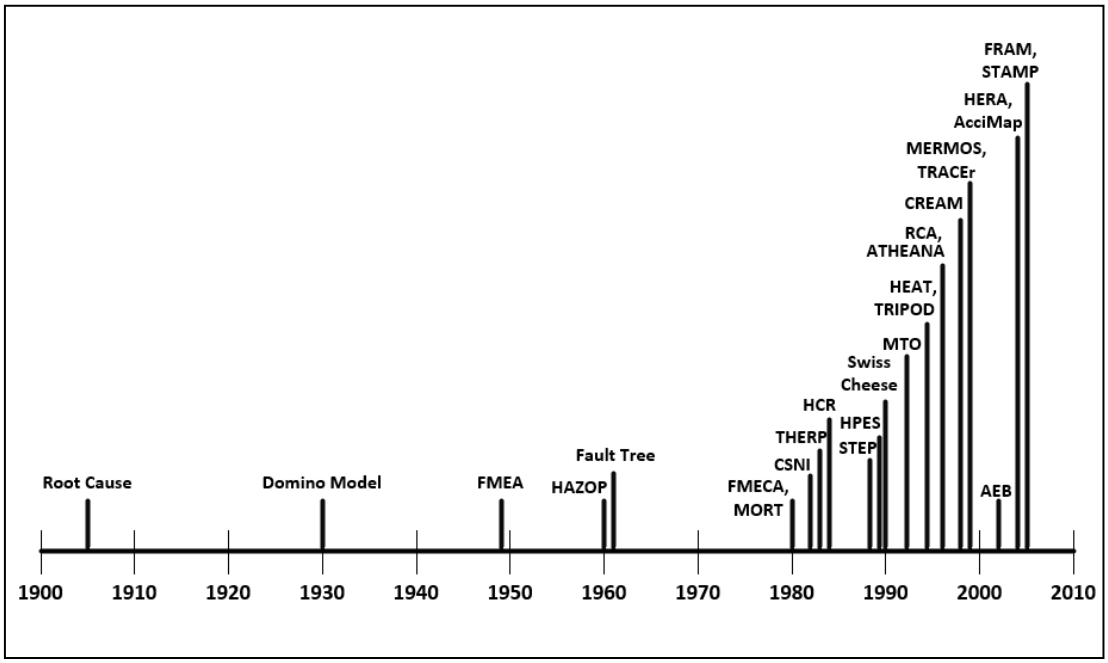


Figure 1. Development of best-known hazard and accident analysis models (adapted from *[6]*).

Over the past decade, the aviation industry also started using systemic methods. For example, [10–12] suggest that such methods are more effective in terms of assessing system complexity and component interactions. Yet, classic methods are still used primarily because of their long presence.

The automotive industry presents an interesting domain in terms of the potential to link modern risk assessment methods with safety standards. For example, ISO 26262 was developed for ensuring the functional safety of the systems in the automotive domain [13]. However, it does not demand a specific method to be included in the risk management process [14]. Nevertheless, whereas in a similar fashion to aviation industry classic methods are popular, some studies such as [15] and [16] have suggested that a systemic method, STPA, can be a better option as it can be applied to a new system design from an early stage to determine the detailed list of functions, failures and mitigation measures, even without having a detailed information of the design.

In the maritime domain, the Formal Safety Assessment introduced by the International Maritime Organization (IMO) has been widely used for the development and use of risk management practices. The IMO FSA framework [17] does not specify the risk methods to be used. Yet, there is a list of approaches (e.g. FTA, FMEA, HAZOP, HAZID) depending on the types of systems and their stage of design or operational implementation/management. With the rapid development of autonomous systems, the necessity to develop more suitable methods that may handle systemic risks and dysfunctional interactions between system components seems essential.

Categories of Analysis Models

Perrow [18] developed a matrix which classifies different domains based on the manageability and coupling of their systems. Underwood and Waterson [4] has then suggested different categories of analysis models that are suitable for the quadrants in the afore-mentioned Perrow-matrix (see Figure 2). These categories are specified as sequential, epidemiological and systemic. Sequential models consist of methods that have a pre-described path and therefore linear correlation between the origin of an accident (the root cause) and the outcome (the effect). The methods such as Domino model, FTA, FMEA and Root Cause Analysis are classified in this category [4]. Epidemiological methods view accidents as a combination of "latent" and "active" failures within the system. Latent conditions link with working practices (e.g. management and organizational culture) that drive the dynamics between good intentions and actual working procedures. The most popular epidemiological models are Swiss Cheese Model, the Human Factor Analysis & Classification System (HFACS), and the ATSB accident investigation model [4]. Finally, systemic models such as STAMP use the application of systems theory and describe accidents as the result of lack of safety constraints to control the scenarios generated due to unsafe component interactions in a system [19].

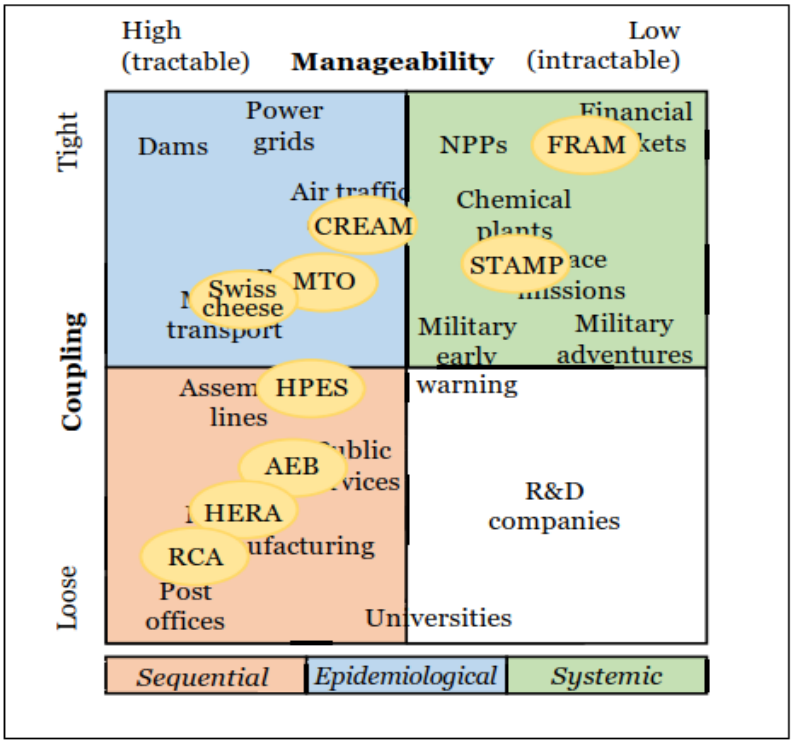


Figure 2. Hazard and accident analysis models categories suitable for different sets of domains in Perrow-matrix (adapted from [4]).

Swot Analysis

A SWOT analysis was then conducted for analyzing the strengths, weaknesses, opportunities and threats of each of the analysis models categories. The SWOT analysis of sequential models, epidemiological models and systemic models are presented in Figure 3, Figure 4 and Figure 5 respectively.

Sequential Models

STRENGTHS

- Simple and less time consuming than other categories as the guidance and taxonomies are available.
- Widely used and well known methods in most of the domains.
- Effective when implemented in simple systems with mainly component failures and human actions (Underwood and Waterson, 2013).
- Identifies the root cause. (Underwood and Waterson, 2014).
- Can include qualitative risk assessment i.e. the estimation of risk probability and consequences (Alexander and Kelly, 2009).

WEAKNESSES

- Identification of component interactions issues in a system is very limited.
- Poor handling of managerial, organizational, human and software components (Underwood and Waterson, 2013). Moreover, the humans and software are treated in the same way as mechanical hardware and are assumed that they fail in the same way.
- Can lead to incorrect / unjustified assigning of blame, which additionally "represents a missed opportunity to learn important lessons about system safety" (Underwood and Waterson, 2014).
- Probabilities can be unrealistic and therefore dangerous.
- Assumes that the component failure modes are independent.

OPPORTUNITIES

- Optimal for the systems with the following characteristics:
  - Simple systems (low level of complexity).
  - Mainly physical components.
  - Loose coupling.
  - High manageability.
  - Systems with the existing databases.
- Optimal in the following conditions:
  - Limited resources (time, money and human).
  - Identification of root cause and assigning blame is required.
  - Suggested/demanded by the regulatory bodies.
  - Qualitative risk assessment is desired.

THREATS

- Not optimal for the systems with the following characteristics:
  - Socio-technical systems with high level of complexity.
  - Systems with the high involvement of humans, organisation, and/or software.
  - Tight coupling.
  - Low manageability.
  - New systems with a limited database.
- Not optimal in the following conditions:
  - Identification of dysfunctional interactions is the priority.
  - A comprehensive list of safety controls improvements to the system is desired.

Figure 3. A SWOT analysis of the sequential models.

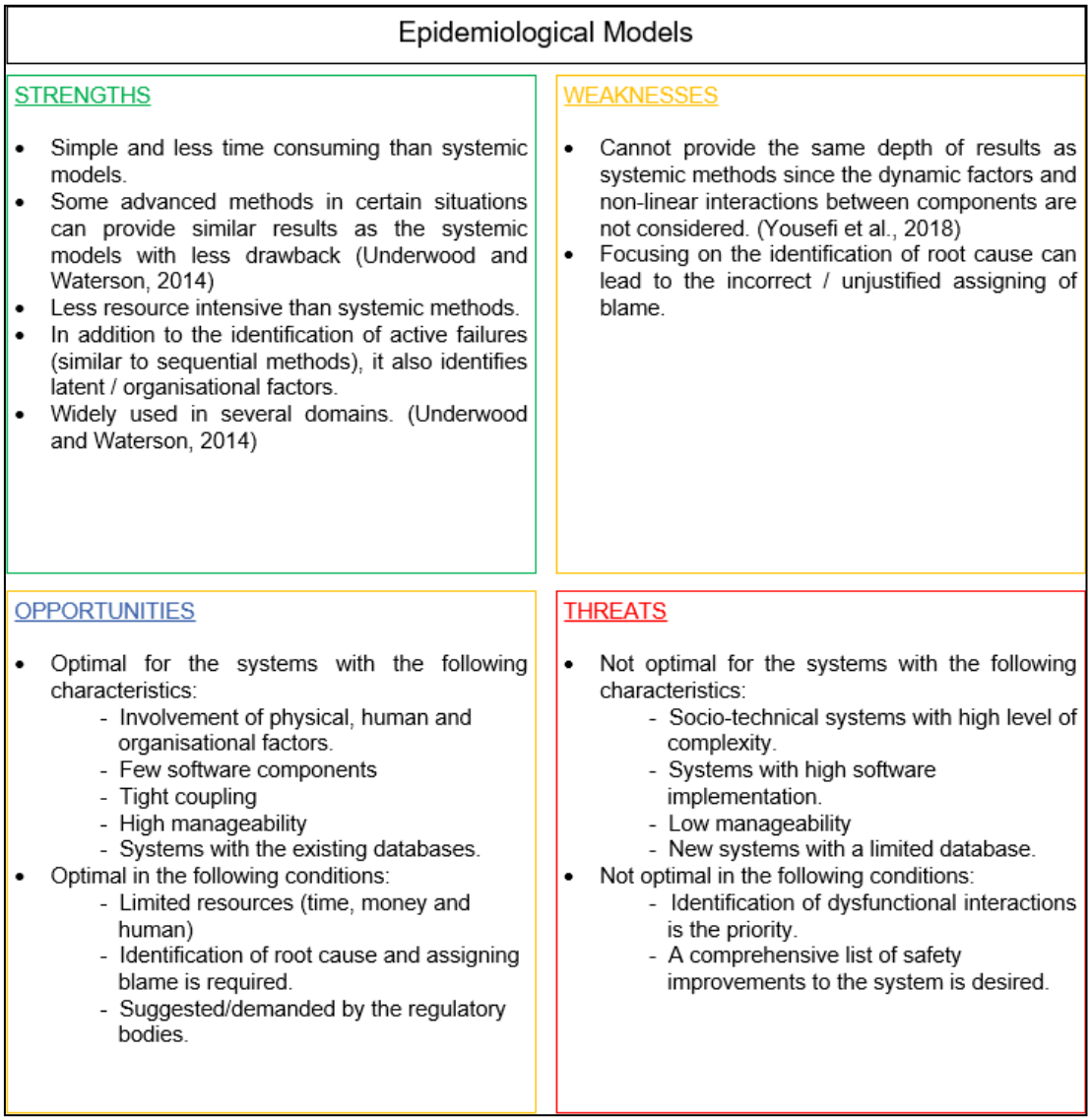


Figure 4. A SWOT analysis of the epidemiological models.

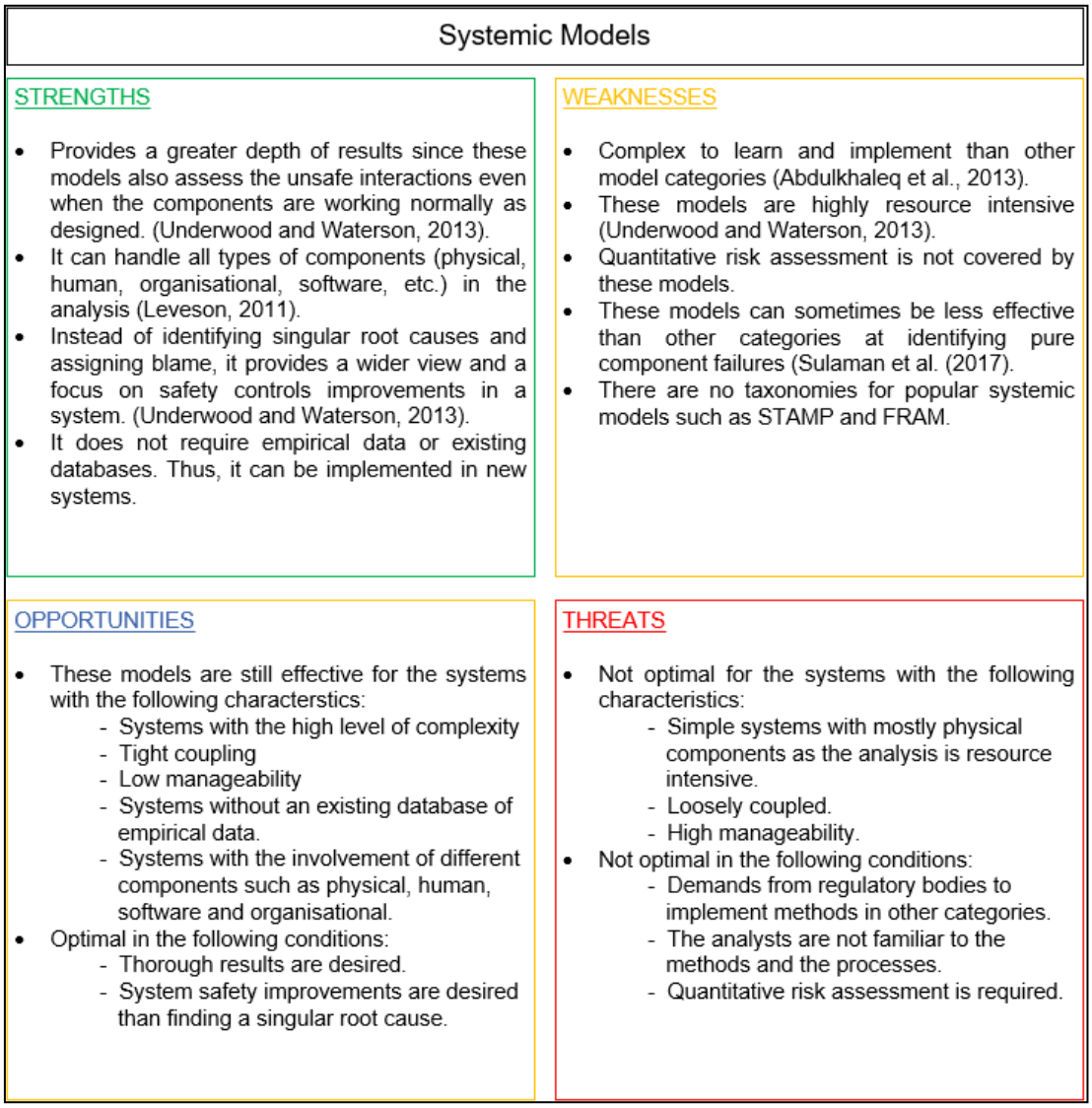


Figure 5. A SWOT analysis of the systemic models.

Discussion

The literature review and the SWOT analysis presented in this paper imply that understanding complex interactions between technologically disruptive systems is an important first step in terms of estimating their potential implementation within the context of managing autonomy related risks and defining risk management systems of relevance based on risk control options. Across multi-modal domains, it becomes obvious that autonomous systems and operations are defined by components and subsystems that are interconnected. In this sense, manageability becomes critical in terms of understanding risks associated with system functionality. Another key point to consider is intractability especially for cases where principles of functioning are unknown, while a high level of detail is essential to understand the dynamic interactions of

sub-systems. Accordingly, the high complexity and interconnectivity of autonomous systems are key factors and should be considered in terms of defining unified approaches and risk management models in autonomous marine domain.

The comparison of the strengths, weaknesses, opportunities and threats of all categories shows that the systemic models are the most suitable models for the systems with tight coupling and low manageability. Furthermore, these models are also effective in systems with the high involvement of different components such as physical, human, organizational and software. Moreover, the systemic models do not require empirical data as these models do not aim to estimate the probability of risk occurrence and consequences. In addition, these models have several other benefits such as providing a wider view and focus on safety control improvements and an assessment of dysfunctional interactions even in normally operating components. As all these features are required in autonomous marine ecosystems, this study shows that the systemic models can be a suitable option to analyses the autonomous systems in marine industry and manage risks from the earliest design phase.

Conclusions

The review presented in this paper suggests that modern risk assessment practices (e.g. FRAM, STAMP) could be a foundation or an optimal choice for the risk assessment of autonomous marine systems especially considering the sub-system complexity and interconnectivity of autonomous ships and their components. Nevertheless, there are also some drawbacks of using such approaches as they require high resources and well-developed educational practices. Considering that the information on autonomous designs and their functionality is still limited, there is a need to re-consider all available methods in a greater detail. Various factors such as (a) the desired level of thoroughness in comparison to resource consumption; (b) the level of detail in the analysis and (c) the format ad content (risk nodes) that may be demanded by each of the categories for the analysis could be considered to justify any future choices.

References

[1] MUNIN. Research in maritime autonomous systems project results and technology potentials. 2016.

[2] Thieme CA, Utne IB, Haugen S. Assessing ship risk model applicability to Marine Autonomous Surface Ships. Ocean Eng 2018;165:140–54. doi:10.1016/j.oceaneng.2018.07.040.

[3] Montewka J, Wróbel K, Heikkilä E, Valdez Banda OA, Goerlandt F, Haugen S.

Challenges, solution proposals and research directions in safety and risk assessment of autonomous shipping. PSAM 14th Probabilistic Saf Assess Manag Conf 2018.

[4] Underwood PJ, Waterson PE. Accident analysis models and methods: guidance for safety professionals. 2013.

[5] Manzur Tirado AM, Brown R, Valdez Banda OA. Risk and Safety management of autonomous systems: a literature review and initial proposals for the maritime industry [online]. 2019.

[6] Hollnagel E. From FRAM (Functional Resonance Accident Model) to FRAM (Functional Resonance Analysis Method) 2008.

[7] Garrett M, Boslaugh SE. Railroad Automation Technology. Encycl. Transp. Soc. Sci. Policy, 2014. doi:10.4135/9781483346526.n396.

[8] Belmonte F, Schön W, Heurley L, Capel R. Interdisciplinary safety analysis of complex socio-technological systems based on the functional resonance accident model: An application to railway trafficsupervision. Reliab Eng Syst Saf 2011. doi:10.1016/j.ress.2010.09.006.

[9] Dong A, Leveson N. Application of Cast and Stpa To Railroad Safety in China. Massachusetts Inst Technol 2012.

[10] Ishimatsu T, Leveson N, Thomas J, Katahira M, Miyamoto Y, Nakao H. Modeling and hazard analysis using STPA. Eur. Sp. Agency, (Special Publ. ESA SP, 2010.

[11] Allison CK, Revell KM, Sears R, Stanton NA. Systems Theoretic Accident Model and Process (STAMP) safety modelling applied to an aircraft rapid decompression event. Saf Sci 2017. doi:10.1016/j.ssci.2017.06.011.

[12] Fleming CH, Spencer M, Thomas J, Leveson N, Wilkinson C. Safety assurance in NextGen and complex transportation systems. Saf Sci 2013. doi:10.1016/j.ssci.2012.12.005.

[13] Czerny BJ, Ambrosio JD, Debouk R. ISO 26262 Functional Safety Draft Intrenational Standard for Road Vehicles: Background, Status and Overview. 2011.

[14] Abdulkhaleq A. Experiences with Applying STPA to Software-Intensive Systems in the Automotive Domain Motivation : STAMP / STPA Application Areas 2013:1–17.

[15] Abdulkhaleq A, Wagner S, Lammering D, Boehmert H, Blueher P. Using STPA in Compliance with ISO 26262 for Developing a Safe Architecture for Fully Automated Vehicles 2017:11–24.

[16] Sabaliauskaite G, Liew LS, Cui J. Integrating Autonomous Vehicle Safety and Security Analysis Using STPA Method and the Six-Step Model. Int J Adv Secur 2018.

[17] International Maritime Organization (IMO). https://edocs.imo.org/Final Documents/English/MSC-MEPC.2-Circ.12-Rev.1 (E).docx. vol. 44. London: 2015.

[18] Perrow C. Normal accidents: Living with high risk technologies. 1999. doi:10.5465/AMR.1985.4278477.

[19] Leveson N. Engineering a safer world: systems thinking applied to safety. The MIT Press; 2016. doi:10.5860/choice.49-6305.

Organizing Committee

Marilia Ramos, PhD

Dr. Marilia Ramos has a PhD in Chemical Engineering from the Federal University of Pernambuco, Brazil. Her expertise is on Risk Analysis and Human Reliability, and she has extensive experience in projects concerning the oil and gas field. She is a postdoctoral research fellow at the Department of Marine Technology, NTNU, and applies risk and reliability analysis to autonomous surface vessels. Her main research interest is on the human-software-hardware interaction in such systems.

Christoph A. Thieme, PhD

Dr. Christoph Thieme obtained his PhD in Marine Technology from NTNU. He has experience with risk analysis and modelling of autonomous marine systems. Currently, he is a postdoctoral research fellow at NTNU in the UNLOCK project, working on risk assessment methods development and applications on autonomous control systems. His main research interest is the software and control system aspects of autonomous systems on the risk level.

Ingrid B. Utne, PhD

Dr. Ingrid Bouwer Utne is a Professor in marine operation and maintenance at Department of Marine Technology, NTNU. Utne is an affiliated Researcher in the Center of Excellence on Autonomous Marine Operations and Systems (NTNU AMOS) where she is managing the research/industry projects UNLOCK and ORCAS. These projects focus on supervisory risk control and bridge the scientific disciplines of risk management and engineering cybernetics aiming to enhance safety and intelligence in autonomous systems.

Ali Mosleh, PhD

Dr. Ali Mosleh is Distinguished University Professor and holder of the Knight Endowed Chair in Engineering at the University of California in Los Angeles (UCLA), where he is also the director of the Institute for the Risk Sciences. He is also honorary professor at several universities in Europe and Asia. He conducts research on methods for probabilistic risk analysis and reliability of complex systems and has made many contributions in diverse fields of theory and application. He was elected to the US National Academy of Engineering in 2010 and is a Fellow of the Society for Risk Analysis, and the American Nuclear Society. Prof. Mosleh is the recipient of many scientific achievement awards.

Organizers and Sponsors

Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

The Department of Marine Technology at NTNU provides world-class education and research for engineering systems in the marine environment. The focus is on methods and techniques for sustainable development and operation of ship technology, fisheries and aquaculture technology, oil and gas extraction at sea, offshore renewable energy, and marine robotics for mapping and monitoring the ocean. The Department hosts an excellent research group working on safety and risk management of marine and maritime systems. The Centre of Excellence Autonomous Marine Operations and Systems (NTNU AMOS) is also located at the Department.

The Norwegian University of Science and Technology in Trondheim (NTNU) is the largest university in Norway.

The B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA

The B. John Garrick Institute for the Risk Sciences has declared its mission to be the advancement and application of the risk sciences to save lives, protect the environment and improve system performance. The purpose of the Garrick Institute is for the research, development, and application of technology for (1) quantifying the risk of the most serious threats to society to better enable their prevention, reduce their likelihood of occurrence or limit their consequences and (2) improving system performance with respect to reliability and safety. The institute is hosted at the Department of Engineering at the University of California Los Angeles (UCLA).

DNV GL

DNV GL is a global quality assurance and risk management company. DNV GL provides classification, technical assurance, software and independent expert advisory services to several industries. Combining technical, digital and operational expertise, risk methodology and in-depth industry knowledge, DNV GL assists its customers in decisions and actions with trust and confidence. With origins stretching back to 1864 and operations in more than 100 countries. DNV GL are dedicated to helping customers make the world safer, smarter and greener.

Rolls Royce Marine

Rolls Royce Marine (RRM, now Kongsberg Maritime) is a leading supplier of offshore and marine energy solutions, deck machinery and automation systems. In addition, RRM provides services related to complex system integration, and vessel design. RRM is a leader in marine ship intelligence, automation and autonomy, testing successfully a fully autonomous ferry transit in Finland in late 2018. RRM is now a part of the Kongsberg Group.

Research Council of Norway

The Research Council of Norway serves as the chief advisory body for the government authorities on research policy issues, and distributes roughly nine billion Norwegian kroners to research and innovation activities each year. The Research Council of Norway co-financed the IWASS workshop through the MAROFF knowledge-building project for industry ORCAS (Project number 280655) and the FRINATEK project UNLOCK (Project number 274441).

IWASS Participants

Organizing committee

Marilia A. Ramos	Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Norway
Christoph A. Thieme	Department of Marine Technology, NTNU, Norway
Ingrid Bouwer Utne	Department of Marine Technology, NTNU, Norway
Ali Mosleh	B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA

Speakers

Adrian Aljornilla	UAS Consulting, USA
Anouck Girard	University of Michigan, USA
Asun Lera St. Clair	DNV GL, Norway
David B. Kaber	North Carolina state university, ISE, USA
Kenneth Titlestad	Sopra Steria, Norway
Matthew Minxiang Hu	Haylion Technologies, China
Nils Haktor Bua	Norwegian Maritime Authority, Norway
Sverre Torben	Rolls Royce Marine, Norway
Tristan Perez	Boeing Research and Technology, Australia

Participants

Andrey Morozov	Technical University, Dresden, Germany
Arne Ulrik Bindingsbø	Equinor, Norway
Asgeir J. Sørensen	NTNU Centre of Excellence for Autonomous Marine Operations and Systems, NTNU, Norway
Bentolhoda Jafary	University of Massachusetts Dartmouth, USA
Bernhard Twomey	Rolls Royce Marine, Norway
Børge Rokseth	Department of Marine Technology, NTNU, Norway
Daniel Metlay	B. John Garrick Institute for the Risk Sciences, University of California, Los Angeles, USA
Edmund Brekke	Department of Engineering Cybernetics, NTNU, Norway
Ingrid Schjølberg	NTNU OCEANS, Norway
Jens Einar Bremnes	Department of Marine Technology, NTNU, Norway
John Andrews	University of Nottingham, United Kingdom
Jon Arne Glomsrud	DNV GL, Norway
Jonas Borg	Volvo Penta, Sweden
Lance Fiondella	University of Massachusetts, USA
Mario Brito	University of Southampton, United Kingdom
Markus Heimdal	Rolls Royce Marine, Norway
Matthew Luckcuck	University of Liverpool, United Kingdom
Nikolaos P. Ventikos	National Technical University of Athens, Greece
Odd Ivar Haugen	DNV GL, Norway
Ole Jakob Mengshoel	Department of Computer Science, NTNU, Norway
Ørnulf Jan Rødseth	SINTEF Ocean, Norway
Osiris Valdez Banda	Aalto University, Finland
Rudolf Mester	Norwegian AI Lab, NTNU, Norway
Salvatore Massaiu	Institute for Energy Technology, Norway

Participants

Simon Blindheim	Department of Engineering Cybernetics, NTNU, Norway
Siri Granum Carson	Department of Philosophy and Religious Studies, NTNU, Norway
Siv Randi Hjørungnes	Rolls Royce Marine, Norway
Stein Haugen	Department of Marine Technology, NTNU, Norway
Stig Ole Johnsen	SINTEF Digital, Norway
Sverre Rothmund	Department of Engineering Cybernetics, NTNU, Norway
Thomas Johansen	Department of Marine Technology, NTNU, Norway
Thomas Porathe	Department of Design, NTNU, Norway
Thor Myklebust	SINTEF Digital, Norway
Tobias Torben	Department of Marine Technology, NTNU, Norway
Tom Mace	UAS Consulting, USA
Tor Arne Johansen	Department of Engineering Cybernetics, NTNU, Norway
Torgeir Moan	Department of Marine Technology, NTNU, Norway
Trine Stene	SINTEF Digital, Norway
Yan-Fu Li	Tsinghua University, China
Yi He	Wuhan University of Technology, China

Published by:
Norwegian University of Science and Technology (NTNU)
Department of Marine Technology
Otto Nielsen Veg 10, 7491 Trondheim, Norway
ISBN: 978-82-691120-2-3

NTNU
Norwegian University of Science and Technology

The B. John Garrick Institute for the Risk Sciences
UCLA ENGINEERING